Probability and Measure

Third Edition

Patrick Billingsley

Probability and Measure

Third Edition

PATRICK BILLINGSLEY

The University of Chicago

Preface

Edward Davenant said he "would have a man knockt in the head that should write anything in Mathematiques that had been written of before." So reports John Aubrey in his *Brief Lives*. What is new here then?

To introduce the idea of measure the book opens with Borel's normal number theorem, proved by calculus alone. and there follow short sections establishing the existence and fundamental properties of probability measures, including Lebesgue measure on the unit interval. For simple random variables—ones with finite range—the expected value is a sum instead of an integral. Measure theory, without integration, therefore suffices for a completely rigorous study of infinite sequences of simple random variables, and this is carried out in the remainder of Chapter 1, which treats laws of large numbers, the optimality of bold play in gambling, Markov chains, large deviations, the law of the iterated logarithm. These developments in their turn motivate the general theory of measure and integration in Chapters 2 and 3.

Measure and integral are used together in Chapters 4 and 5 for the study of random sums, the Poisson process, convergence of measures, characteristic functions, central limit theory. Chapter 6 begins with derivatives according to Lebesgue and Radon–Nikodym—a return to measure theory—then applies them to conditional expected values and martingales. Chapter 7 treats such topics in the theory of stochastic processes as Kolmogorov's existence theorem and separability, all illustrated by Brownian motion.

What is new, then, is the alternation of probability and measure, probability motivating measure theory and measure theory generating further probability. The book presupposes a knowledge of combinatorial and discrete probability, of rigorous calculus, in particular infinite series, and of elementary set theory. Chapters 1 through 4 are designed to be taken up in sequence. Apart from starred sections and some examples, Chapters 5, 6, and 7 are independent of one another; they can be read in any order.

My goal has been to write a book I would myself have liked when I first took up the subject, and the needs of students have been given precedence over the requirements of logical economy. For instance, Kolmogorov's exis-

tence theorem appears not in the first chapter but in the last, stochastic processes needed earlier having been constructed by special arguments which, although technically redundant, motivate the general result. And the general result is, in the last chapter, given two proofs at that. It is instructive, I think, to see the show in rehearsal as well as in performance.

The Third Edition. The main changes in this edition are two For the theory of Hausdorff measures in Section 19 I have substituted an account of L^p spaces, with applications to statistics. And for the queueing theory in Section 24 I have substituted an introduction to ergodic theory, with applications to continued fractions and Diophantine approximation. These sections now fit better with the rest of the book, and they illustrate again the connections probability theory has with applied mathematics on the one hand and with pure mathematics on the other.

For suggestions that have led to improvements in the new edition, I thank Raj Bahadur, Walter Philipp, Michael Wichura, and Wing Wong, as well as the many readers who have sent their comments.

Envoy. I said in the preface to the second edition that there would not be a third, and yet here it is. There will not be a fourth. It has been a very agreeable labor, writing these successive editions of my contribution to the river of mathematics. And although the contribution is small, the river is great: After ages of good service done to those who people its banks, as Joseph Conrad said of the Thames, it spreads out "in the tranquil dignity of a waterway leading to the uttermost ends of the earth."

PATRICK BILLINGSLEY

Chicago, Illinois
December 1994

Contents

---

*Stars indicate topics that may be omitted on a first reading.

Probability and Measure

CHAPTER 1

Probability

SECTION 1. BOREL'S NORMAL NUMBER THEOREM

Although sufficient for the development of many interesting topics in mathematical probability, the theory of discrete probability spaces[†] does not go far enough for the rigorous treatment of problems of two kinds: those involving an infinitely repeated operation, as an infinite sequence of tosses of a coin, and those involving an infinitely fine operation, as the random drawing of a point from a segment. A mathematically complete development of probability, based on the theory of measure, puts these two classes of problem on the same footing, and as an introduction to measure-theoretic probability it is the purpose of the present section to show by example why this should be so.

The Unit Interval

The project is to construct simultaneously a model for the random drawing of a point from a segment and a model for an infinite sequence of tosses of a coin. The notions of independence and expected value, familiar in the discrete theory, will have analogues here, and some of the terminology of the discrete theory will be used in an informal way to motivate the development. The formal mathematics, however, which involves only such notions as the length of an interval and the Riemann integral of a step function, will be entirely rigorous. All the ideas will reappear later in more general form.

Let Ω denote the unit interval $(0, 1]$; to be definite, take intervals open on the left and closed on the right. Let ω denote the generic point of Ω. Denote the length of an interval $I = (a, b]$ by $|I|$:

$$(1.1) \qquad |I| = |(a, b]| = b - a.$$

[†]For the discrete theory, presupposed here, see for example the first half of Volume I of FELLER. (Names in capital letters refer to the bibliography on p. 581)

If

$$(1.2) \qquad A = \bigcup_{i=1}^{n} I_i = \bigcup_{i=1}^{n} (a_i, b_i],$$

where the intervals $I_i = (a_i, b_i]$ are disjoint [A3][†] and are contained in Ω, assign to A the probability

$$(1.3) \qquad P(A) = \sum_{i=1}^{n} |I_i| = \sum_{i=1}^{n} (b_i - a_i).$$

It is important to understand that in this section $P(A)$ is defined only if A is a finite disjoint union of subintervals of $(0, 1]$—never for sets A of any other kind.

If A and B are two such finite disjoint unions of intervals, and if A and B are disjoint, then $A \cup B$ is a finite disjoint union of intervals and

$$(1.4) \qquad P(A \cup B) = P(A) + P(B).$$

This relation, which is certainly obvious intuitively, is a consequence of the additivity of the Riemann integral:

$$(1.5) \qquad \int_0^1 (f(\omega) + g(\omega))\, d\omega = \int_0^1 f(\omega)\, d\omega + \int_0^1 g(\omega)\, d\omega.$$

If $f(\omega)$ is a step function taking value c_j in the interval $(x_{j-1}, x_j]$, where $0 = x_0 < x_1 < \cdots < x_k = 1$, then its integral in the sense of Riemann has the value

$$(1.6) \qquad \int_0^1 f(\omega)\, d\omega = \sum_{j=1}^{k} c_j (x_j - x_{j-1}).$$

If $f = I_A$ and $g = I_B$ are the indicators [A5] of A and B, then (1.4) follows from (1.5) and (1.6), provided A and B are disjoint. This also shows that the definition (1.3) is unambiguous—note that A will have many representations of the form (1.2) because $(a, b] \cup (b, c] = (a, c]$. Later these facts will be derived anew from the general theory of Lebesgue integration.[‡]

According to the usual models, if a radioactive substance has emitted a single α-particle during a unit interval of time, or if a single telephone call has arrived at an exchange during a unit interval of time, then the instant at which the emission or the arrival occurred is random in the sense that it lies in (1.2) with probability (1.3). Thus (1.3) is the starting place for the

[†]A notation [An] refers to paragraph n of the appendix beginning on p. 536; this is a collection of mathematical definitions and facts required in the text.
[‡]Passages in small type concern side issues and technical matters, but their contents are sometimes required later.

description of a point drawn at random from the unit interval: Ω is regarded as a sample space, and the set (1.2) is identified with the event that the random point lies in it.

The definition (1.3) is also the starting point for a mathematical representation of an infinite sequence of tosses of a coin. With each ω associate its nonterminating dyadic expansion

$$(1.7) \qquad \omega = \sum_{n=1}^{\infty} \frac{d_n(\omega)}{2^n} = .d_1(\omega)d_2(\omega)\ldots,$$

each $d_n(\omega)$ being 0 or 1 [A31]. Thus

$$(1.8) \qquad \left(d_1(\omega), d_2(\omega), \ldots\right)$$

is the sequence of binary digits in the expansion of ω. For definiteness, a point such as $\frac{1}{2} = .1000\ldots = .0111\ldots$, which has two expansions, takes the nonterminating one; 1 takes the expansion $.111\ldots$.

Graph of $d_1(\omega)$ Graph of $d_2(\omega)$
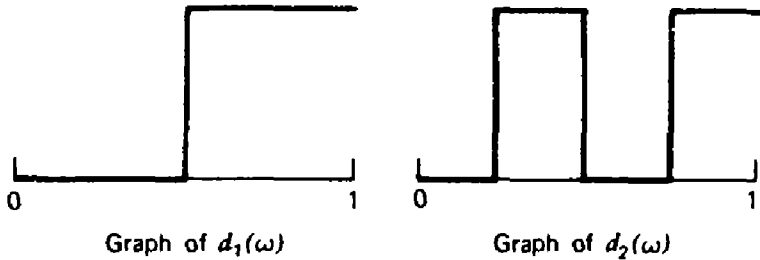
Imagine now a coin with faces labeled 1 and 0 instead of the usual heads and tails. If ω is drawn at random, then (1.8) behaves as if it resulted from an infinite sequence of tosses of a coin. To see this, consider first the set of ω for which $d_i(\omega) = u_i$ for $i = 1, \ldots, n$, where $u_1, \ldots, u_n$ is a sequence of 0's and 1's. Such an ω satisfies

$$\sum_{i=1}^{n} \frac{u_i}{2^i} < \omega \le \sum_{i=1}^{n} \frac{u_i}{2^i} + \sum_{i=n+1}^{\infty} \frac{1}{2^i},$$

where the extreme values of ω correspond to the case $d_i(\omega) = 0$ for $i > n$ and the case $d_i(\omega) = 1$ for $i > n$. The second case can be achieved, but since the binary expansions represented by the $d_i(\omega)$ are nonterminating—do not end in 0's—the first cannot, and ω must actually exceed $\sum_{i=1}^{n} u_i / 2^i$. Thus

$$(1.9) \qquad \left[\omega : d_i(\omega) = u_i, i = 1, \ldots, n\right] = \left(\sum_{i=1}^{n} \frac{u_i}{2^i}, \sum_{i=1}^{n} \frac{u_i}{2^i} + \frac{1}{2^n}\right].$$

The interval here is open on the left and closed on the right precisely because the expansion (1.7) is the nonterminating one. In the model for coin tossing the set (1.9) represents the event that the first n tosses give the outcomes $u_1, \ldots, u_n$ in sequence. By (1.3) and (1.9),

$$(1.10) \qquad P\{\omega: d_i(\omega) = u_i, i = 1, \ldots, n\} = \frac{1}{2^n},$$

which is what probabilistic intuition requires.

Decompositions by dyadic intervals
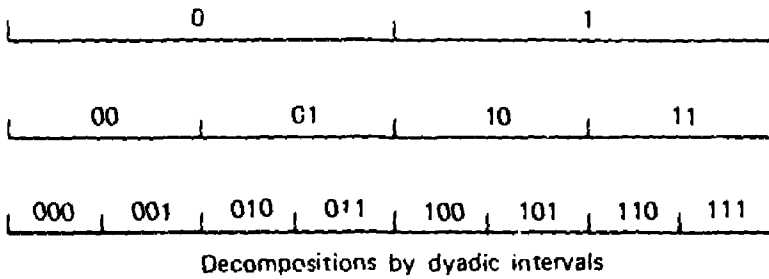
The intervals (1.9) are called *dyadic* intervals, the endpoints being adjacent dyadic rationals $k/2^n$ and $(k + 1)/2^n$ with the same denominator, and n is the *rank* or *order* of the interval. For each n the 2^n dyadic intervals of rank n decompose or partition the unit interval. In the passage from the partition for n to that for $n + 1$, each interval (1.9) is split into two parts of equal length, a left half on which $d_{n+1}(\omega)$ is 0 and a right half on which $d_{n+1}(\omega)$ is 1. For $u = 0$ and for $u = 1$, the set $\{\omega: d_{n+1}(\omega) = u\}$ is thus a disjoint union of 2^n intervals of length $1/2^{n+1}$ and hence has probability $\frac{1}{2}$: $P\{\omega: d_n(\omega) = u\} = \frac{1}{2}$ for all n.

Note that $d_i(\omega)$ is constant over each dyadic interval of rank i and that for $n > i$ each dyadic interval of rank n is entirely contained in a single dyadic interval of rank i. Therefore, $d_i(\omega)$ is constant over each dyadic interval of rank n if $i \leq n$.

The probabilities of various familiar events can be written down immediately. The sum $\sum_{i=1}^{n} d_i(\omega)$ is the number of 1's among $d_1(\omega), \ldots, d_n(\omega)$, to be thought of as the number of heads in n tosses of a fair coin. The usual binomial formula is

$$(1.11) \qquad P\left[\omega: \sum_{i=1}^{n} d_i(\omega) = k\right] = \binom{n}{k} \frac{1}{2^n}, \qquad 0 \leq k \leq n.$$

This follows from the definitions: The set on the left in (1.11) is the union of those intervals (1.9) corresponding to sequences $u_1, \ldots, u_n$ containing k 1's and $n - k$ 0's; each such interval has length $1/2^n$ by (1.10) and there are $\binom{n}{k}$ of them, and so (1.11) follows from (1.3).

The functions $d_n(\omega)$ can be looked at in two ways. Fixing n and letting ω vary gives a real function $d_n = d_n(\cdot)$ on the unit interval. Fixing ω and letting n vary gives the sequence (1.8) of 0's and 1's. The probabilities (1.10) and (1.11) involve only finitely many of the components $d_i(\omega)$. The interest here, however, will center mainly on properties of the entire sequence (1.8). It will be seen that the mathematical properties of this sequence mirror the properties to be expected of a coin-tossing process that continues forever.

As the expansion (1.7) is the nonterminating one, there is the defect that for no ω is (1.8) the sequence $(1,0,0,0,\ldots)$, for example. It seems clear that the chance should be 0 for the coin to turn up heads on the first toss and tails forever after, so that the absence of $(1,0,0,0,\ldots)$—or of any other single sequence—should not matter. See on this point the additional remarks immediately preceding Theorem 1.2.

The Weak Law of Large Numbers

In studying the connection with coin tossing it is instructive to begin with a result that can, in fact, be treated within the framework of discrete probability, namely, the *weak law of large numbers*:

Theorem 1.1. *For each ϵ,[†]*

$$(1.12) \qquad \lim_{n \to \infty} P\left[\omega: \left|\frac{1}{n}\sum_{i=1}^{n} d_i(\omega) - \frac{1}{2}\right| \geq \epsilon\right] = 0.$$

Interpreted probabilistically, (1.12) says that if n is large, then there is small probability that the fraction or relative frequency of heads in n tosses will deviate much from $\frac{1}{2}$, an idea lying at the base of the frequency conception of probability. As a statement about the structure of the real numbers, (1.12) is also interesting arithmetically.

Since $d_i(\omega)$ is constant over each dyadic interval of rank n if $i \leq n$, the sum $\sum_{i=1}^{n} d_i(\omega)$ is also constant over each dyadic interval of rank n. The set in (1.12) is therefore the union of certain of the intervals (1.9), and so its probability is well defined by (1.3).

With the Riemann integral in the role of expected value, the usual application of Chevyshev's inequality will lead to a proof of (1.12). The argument becomes simpler if the $d_n(\omega)$ are replaced by the *Rademacher functions*,

$$(1.13) \qquad r_n(\omega) = 2d_n(\omega) - 1 = \begin{cases} +1 & \text{if } d_n(\omega) = 1, \\ -1 & \text{if } d_n(\omega) = 0. \end{cases}$$

[†]The standard ϵ and δ of analysis will always be understood to be positive.

Graph of $r_1(\omega)$ Graph of $r_2(\omega)$
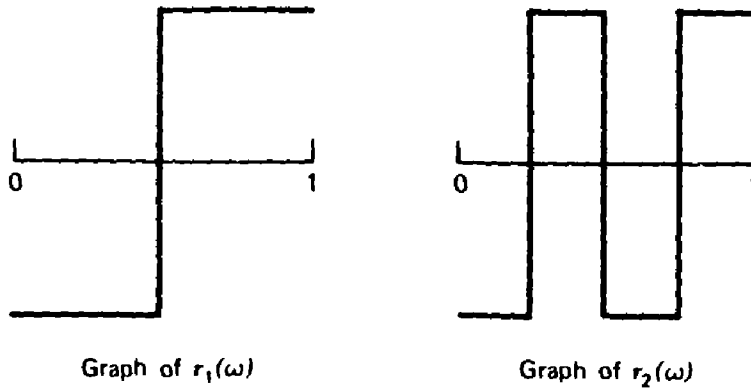
Consider the partial sums

$$(1.14) \qquad\qquad s_n(\omega) = \sum_{i=1}^{n} r_i(\omega).$$

Since $\sum_{i=1}^{n} d_i(\omega) = (s_n(\omega) + n)/2$, (1.12) with $\epsilon/2$ in place of ϵ is the same thing as

$$(1.15) \qquad\qquad \lim_{n \to \infty} P\left[\omega : \left|\frac{1}{n} s_n(\omega)\right| \geq \epsilon\right] = 0.$$

This is the form in which the theorem will be proved.

The Rademacher functions have themselves a direct probabilistic meaning. If a coin is tossed successively, and if a particle starting from the origin performs a random walk on the real line by successively moving one unit in the positive or negative direction according as the coin falls heads or tails, then $r_i(\omega)$ represents the distance it moves on the ith step and $s_n(\omega)$ represents its position after n steps. There is also the gambling interpretation: If a gambler bets one dollar, say, on each toss of the coin, $r_i(\omega)$ represents his gain or loss on the ith play and $s_n(\omega)$ represents his gain or loss in n plays.

Each dyadic interval of rank $i - 1$ splits into two dyadic intervals of rank i; $r_i(\omega)$ has value -1 on one of these and value $+1$ on the other. Thus $r_i(\omega)$ is -1 on a set of intervals of total length $\frac{1}{2}$ and $+1$ on a set of total length $\frac{1}{2}$. Hence $\int_0^1 r_i(\omega)\, d\omega = 0$ by (1.6), and

$$(1.16) \qquad\qquad \int_0^1 s_n(\omega)\, d\omega = 0$$

by (1.5). If the integral is viewed as an expected value, then (1.16) says that the mean position after n steps of a random walk is 0.

Suppose that $i < j$. On a dyadic interval of rank $j - 1$, $r_i(\omega)$ is constant and $r_j(\omega)$ has value -1 on the left half and $+1$ on the right. The product

$r_i(\omega)r_j(\omega)$ therefore integrates to 0 over each of the dyadic intervals of rank $j-1$, and so

(1.17) $$\int_0^1 r_i(\omega)r_j(\omega)\,d\omega = 0, \qquad i \neq j.$$

This corresponds to the fact that independent random variables are uncorrelated. Since $r_i^2(\omega) = 1$, expanding the square of the sum (1.14) shows that

(1.18) $$\int_0^1 s_n^2(\omega)\,d\omega = n.$$

This corresponds to the fact that the variances of independent random variables add. Of course (1.16), (1.17), and (1.18) stand on their own, in no way depend on any probabilistic interpretation.

Applying Chebyshev's inequality in a formal way to the probability in (1.15) now leads to

(1.19) $$P\big[\omega\colon |s_n(\omega)| \geq n\epsilon\big] \leq \frac{1}{n^2\epsilon^2}\int_0^1 s_n^2(\omega)\,d\omega = \frac{1}{n\epsilon^2}.$$

The following lemma justifies the inequality.

Let f be a step function as in (1.6): $f(\omega) = c_j$ for $\omega \in (x_{j-1}, x_j]$, where $0 = x_0 < \cdots < x_k = 1$.

Lemma. *If f is a nonnegative step function, then $[\omega\colon f(\omega) \geq \alpha]$ is for $\alpha > 0$ a finite union of intervals and*

(1.20) $$P\big[\omega\colon f(\omega) \geq \alpha\big] \leq \frac{1}{\alpha}\int_0^1 f(\omega)\,d\omega.$$

The shaded region has area $\alpha P[\omega\colon f(\omega) \geq \alpha]$.
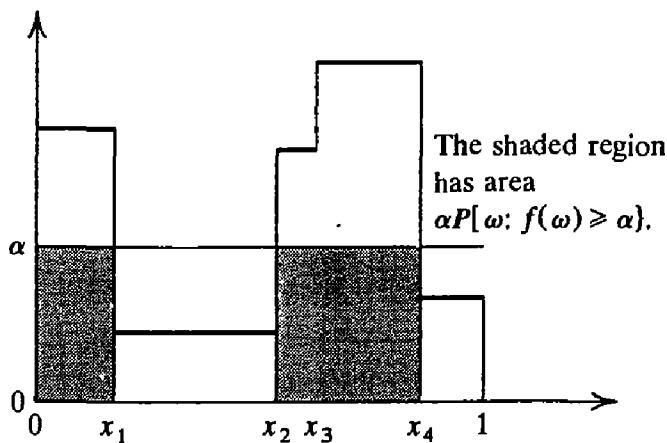
PROOF. The set in question is the union of the intervals $(x_{j-1}, x_j]$ for which $c_j \geq \alpha$. If Σ' denotes summation over those j satisfying $c_j \geq \alpha$, then $P[\omega\colon f(\omega) \geq \alpha] = \Sigma'(x_j - x_{j-1})$ by the definition (1.3). On the other hand,

since the c_j are all nonnegative by hypothesis, (1.6) gives

$$\int_0^1 f(\omega)\, d\omega = \sum_{j=1}^k c_j(x_j - x_{j-1}) \geq {\sum}' c_j(x_j - x_{j-1})$$

$$\geq {\sum}' \alpha(x_j - x_{j-1}).$$

Hence (1.20).   ■

Taking $\alpha = n^2\epsilon^2$ and $f(\omega) = s_n^2(\omega)$ in (1.20) gives (1.19). Clearly, (1.19) implies (1.15), and as already observed, this in turn implies (1.12).

The Strong Law of Large Numbers

It is possible with a minimum of technical apparatus to prove a stronger result that cannot even be formulated in the discrete theory of probability. Consider the set

$$(1.21) \qquad\qquad N = \left[\omega: \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^n d_i(\omega) = \frac{1}{2}\right]$$

consisting of those ω for which the asymptotic relative frequency* of 1 in the sequence (1.8) is $\frac{1}{2}$. The points in (1.21) are called *normal numbers*. The idea is to show that a real number ω drawn at random from the unit interval is "practically certain" to be normal, or that there is "practical certainty" that 1 occurs in the sequence (1.8) of tosses with asymptotic relative frequency $\frac{1}{2}$. It is impossible at this stage to prove that $P(N) = 1$, because N is not a finite union of intervals and so has been assigned no probability. But the notion of "practical certainty" can be formalized in the following way.

Define a subset A of Ω to be *negligible*[†] if for each positive ϵ there exists a finite or countable[‡] collection $I_1, I_2, \ldots$ of intervals (they may overlap) satisfying

$$(1.22) \qquad\qquad A \subset \bigcup_k I_k$$

and

$$(1.23) \qquad\qquad \sum_k |I_k| < \epsilon.$$

A negligible set is one that can be covered by intervals the total sum of whose lengths can be made arbitrarily small. If $P(A)$ is assigned to such an

---

*The *frequency* of 1 (the number of occurrences of it) among $d_1(\omega), \ldots, d_n(\omega)$ is $\sum_{i=1}^n d_i(\omega)$, the *relative* frequency is $n^{-1}\sum_{i=1}^n d_i(\omega)$, and the *asymptotic* relative frequency is the limit in (1.21).
[†]The term *negligible* is introduced for the purposes of this section only. The negligible sets will reappear later as the sets of Lebesgue measure 0.
[‡]*Countably infinite* is unambiguous. *Countable* will mean finite or countably infinite, although it will sometimes for emphasis be expanded as here to *finite or countable*.

A in any reasonable way, then for the I_k of (1.22) and (1.23) it ought to be true that $P(A) \leq \Sigma_k P(I_k) = \Sigma_k |I_k| < \epsilon$, and hence $P(A)$ ought to be 0. Even without any assignment of probability at all, the definition of negligibility can serve as it stands as an explication of "practical impossibility" and "practical certainty": Regard it as practically impossible that the random ω will lie in A if A is negligible, and regard it as practically certain that ω will lie in A if its complement A^c [A1] is negligible.

Although the fact plays no role in the next proof, for an understanding of negligibility observe first that *a finite or countable union of negligible sets is negligible.* Indeed, suppose that $A_1, A_2, \ldots$ are negligible. Given ϵ, for each n choose intervals $I_{n1}, I_{n2}, \ldots$ such that $A_n \subset \bigcup_k I_{nk}$ and $\Sigma_k |I_{nk}| < \epsilon/2^n$. All the intervals I_{nk} taken together form a countable collection covering $\bigcup_n A_n$, and their lengths add to $\Sigma_n \Sigma_k |I_{nk}| < \Sigma_n \epsilon/2^n = \epsilon$. Therefore, $\bigcup_n A_n$ is negligible.

A set consisting of a single point is clearly negligible, and so every countable set is also negligible. The rationals for example form a negligible set. In the coin-tossing model, a single point of the unit interval has the role of a single sequence of 0's and 1's, or of a single sequence of heads and tails. It corresponds with intuition that it should be "practically impossible" to toss a coin infinitely often and realize any one particular infinite sequence set down in advance. It is for this reason not a real shortcoming of the model that for no ω is (1.8) the sequence $(1, 0, 0, 0, \ldots)$. In fact, since a countable set is negligible, it is not a shortcoming that (1.8) is never one of the countably many sequences that end in 0's.

Theorem 1.2.   *The set of normal numbers has negligible complement.*

This is *Borel's normal number theorem,*[†] a special case of the *strong law of large numbers.* Like Theorem 1.1, it is of arithmetic as well as probabilistic interest.

The set N^c is not countable: Consider a point ω for which $(d_1(\omega), d_2(\omega), \ldots) = (1, 1, u_3, 1, 1, u_6, \ldots)$—that is, a point for which $d_i(\omega) = 1$ unless i is a multiple of 3. Since $n^{-1}\Sigma_{i=1}^{n} d_i(\omega) \geq \frac{2}{3}$, such a point cannot be normal. But there are uncountably many such points, one for each infinite sequence $(u_3, u_6, \ldots)$ of 0's and 1's. Thus one cannot prove N^c negligible by proving it countable, and a deeper argument is required.

PROOF OF THEOREM 1.2.   Clearly (1.21) and

$$(1.24) \qquad\qquad N = \left[\omega\colon \lim_{n \to \infty} \frac{1}{n} s_n(\omega) = 0 \right]$$

†Émile Borel: Sur les probabilités dénombrables et leurs applications arithmétiques, *Circ. Mat d. Palermo,* **29** (1909), 247–271. See DUDLEY for excellent historical notes on analysis and probability.

define the same set (see (1.14)). To prove N^c negligible requires constructing coverings that satisfy (1.22) and (1.23) for $A = N^c$. The construction makes use of the inequality

$$(1.25) \qquad P\big[\omega: |s_n(\omega)| \ge n\epsilon\big] \le \frac{1}{n^4\epsilon^4} \int_0^1 s_n^4(\omega)\, d\omega.$$

This follows by the same argument that leads to the inequality in (1.19)—it is only necessary to take $f(\omega) = s_n^4(\omega)$ and $\alpha = n^4\epsilon^4$ in (1.20). As the integral in (1.25) will be shown to have order n^2, the inequality is stronger than (1.19).

The integrand on the right in (1.25) is

$$(1.26) \qquad s_n^4(\omega) = \sum r_\alpha(\omega) r_\beta(\omega) r_\gamma(\omega) r_\delta(\omega),$$

where the four indices range independently from 1 to n. Depending on how the indices match up, each term in this sum reduces to one of the following five forms, where in each case the indices are now *distinct*:

$$(1.27) \qquad \begin{cases} r_i^4(\omega) = 1, \\ r_i^2(\omega) r_j^2(\omega) = 1, \\ r_i^2(\omega) r_j(\omega) r_k(\omega) = r_j(\omega) r_k(\omega), \\ r_i^3(\omega) r_j(\omega) = r_i(\omega) r_j(\omega), \\ r_i(\omega) r_j(\omega) r_k(\omega) r_l(\omega). \end{cases}$$

If, for example, k exceeds i, j, and l, then the last product in (1.27) integrates to 0 over each dyadic interval of rank $k - 1$, because $r_i(\omega) r_j(\omega) r_l(\omega)$ is constant there, while $r_k(\omega)$ is -1 on the left half and $+1$ on the right. Adding over the dyadic intervals of rank $k - 1$ gives

$$\int_0^1 r_i(\omega) r_j(\omega) r_k(\omega) r_l(\omega)\, d\omega = 0.$$

This holds whenever the four indices are distinct. From this and (1.17) it follows that the last three forms in (1.27) integrate to 0 over the unit interval; of course, the first two forms integrate to 1.

The number of occurrences in the sum (1.26) of the first form in (1.27) is n. The number of occurrences of the second form is $3n(n - 1)$, because there are n choices for the α in (1.26), three ways to match it with β, γ, or δ, and $n - 1$ choices for the value common to the remaining two indices. A term-by-term integration of (1.26) therefore gives

$$(1.28) \qquad \int_0^1 s_n^4(\omega)\, d\omega = n + 3n(n - 1) \le 3n^2,$$

and it follows by (1.25) that

$$(1.29) \qquad P\left[\omega: \left|\frac{1}{n}s_n(\omega)\right| \geq \epsilon\right] \leq \frac{3}{n^2\epsilon^4}.$$

Fix a positive sequence $\{\epsilon_n\}$ going to 0 slowly enough that the series $\Sigma_n \epsilon_n^{-4} n^{-2}$ converges (take $\epsilon_n = n^{-1/8}$, for example). If $A_n = [\omega: |n^{-1}s_n(\omega)| \geq \epsilon_n]$, then $P(A_n) \leq 3\epsilon_n^{-4}n^{-2}$ by (1.29), and so $\Sigma_n P(A_n) < \infty$.

If, for some m, ω lies in A_n^c for all n greater than or equal to m, then $|n^{-1}s_n(\omega)| < \epsilon_n$ for $n \geq m$, and it follows that ω is normal because $\epsilon_n \to 0$ (see (1.24)). In other words, for each m, $\bigcap_{n=m}^{\infty} A_n^c \subset N$, which is the same thing as $N^c \subset \bigcup_{n=m}^{\infty} A_n$. This last relation leads to the required covering: Given ϵ, choose m so that $\Sigma_{n=m}^{\infty} P(A_n) < \epsilon$. Now A_n is a finite disjoint union $\bigcup_k I_{nk}$ of intervals with $\Sigma_k |I_{nk}| = P(A_n)$, and therefore $\bigcup_{n=m}^{\infty} A_n$ is a countable union $\bigcup_{n=m}^{\infty} \bigcup_k I_{nk}$ of intervals (not disjoint, but that does not matter) with $\Sigma_{n=m}^{\infty} \Sigma_k |I_{nk}| = \Sigma_{n=m}^{\infty} P(A_n) < \epsilon$. The intervals I_{nk} ($n \geq m$, $k \geq 1$) provide a covering of N^c of the kind the definition of negligibility calls for.   ∎

Strong Law Versus Weak

Theorem 1.2 is stronger than Theorem 1.1. A consideration of the forms of the two propositions will show that the strong law goes far beyond the weak law.

For each n let $f_n(\omega)$ be a step function on the unit interval, and consider the relation

$$(1.30) \qquad \lim_{n \to \infty} P\left[\omega: |f_n(\omega)| \geq \epsilon\right] = 0$$

together with the set

$$(1.31) \qquad \left[\omega: \lim_{n \to \infty} f_n(\omega) = 0\right].$$

If $f_n(\omega) = n^{-1}s_n(\omega)$, then (1.30) reduces to the weak law (1.15), and (1.31) coincides with the set (1.24) of normal numbers. According to a general result proved below (Theorem 5.2(ii)), whatever the step functions $f_n(\omega)$ may be, if the set (1.31) has negligible complement, then (1.30) holds for each positive ϵ. For this reason, a proof of Theorem 1.2 is automatically a proof of Theorem 1.1.

The converse, however, fails: There exist step functions $f_n(\omega)$ that satisfy (1.30) for each positive ϵ but for which (1.31) fails to have negligible complement (Example 5.4). For this reason, a proof of Theorem 1.1 is not automatically a proof of Theorem 1.2; the latter lies deeper and its proof is correspondingly more complex.

Length

According to Theorem 1.2, the complement N^c of the set of normal numbers is negligible. What if N itself were negligible? It would then follow that $(0, 1] = N \cup N^c$ was negligible as well, which would disqualify negligibility as an explication of "practical impossibility," as a stand-in for "probability zero." The proof below of the "obvious" fact that an interval of positive

length is not negligible (Theorem 1.3(ii)), while simple enough, does involve the most fundamental properties of the real number system.

Consider an interval $I = (a, b]$ of length $|I| = b - a$; see (1.1). Consider also a finite or infinite sequence of intervals $I_k = (a_k, b_k]$. While each of these intervals is bounded, they need not be subintervals of $(0, 1]$.

Theorem 1.3. (i) *If* $\bigcup_k I_k \subset I$, *and the* I_k *are disjoint, then* $\sum_k |I_k| \leq |I|$.

(ii) *If* $I \subset \bigcup_k I_k$ (*the* I_k *need not be disjoint*), *then* $|I| \leq \sum_k |I_k|$.

(iii) *If* $I = \bigcup_k I_k$, *and the* I_k *are disjoint, then* $|I| = \sum_k |I_k|$.

PROOF.   Of course (iii) follows from (i) and (ii).

PROOF OF (i): *Finite case.* Suppose there are n intervals. The result being obvious for $n = 1$, assume that it holds for $n - 1$. If a_n is the largest among $a_1, \ldots, a_n$ (this is just a matter of notation), then $\bigcup_{k=1}^{n-1}(a_k, b_k] \subset (a, a_n]$, so that $\sum_{k=1}^{n-1}(b_k - a_k) \leq a_n - a$ by the induction hypothesis, and hence $\sum_{k=1}^{n}(b_k - a_k) \leq (a_n - a) + (b_n - a_n) \leq b - a$.

Infinite case. If there are infinitely many intervals, each finite subcollection satisfies the hypotheses of (i), and so $\sum_{k=1}^{n}(b_k - a_k) \leq b - a$ by the finite case. But as n is arbitrary, the result follows.

PROOF OF (ii): *Finite case.*   Assume that the result holds for the case of $n - 1$ intervals and that $(a, b] \subset \bigcup_{k=1}^{n}(a_k, b_k]$. Suppose that $a_n < b \leq b_n$ (notation again). If $a_n \leq a$, the result is obvious. Otherwise, $(a, a_n] \subset \bigcup_{k=1}^{n-1}(a_k, b_k]$, so that $\sum_{k=1}^{n-1}(b_k - a_k) \geq a_n - a$ by the induction hypothesis and hence $\sum_{k=1}^{n}(b_k - a_k) \geq (a_n - a) + (b_n - a_n) \geq b - a$. The finite case thus follows by induction.

Infinite case. Suppose that $(a, b] \subset \bigcup_{k=1}^{\infty}(a_k, b_k]$. If $0 < \epsilon < b - a$, the open intervals $(a_k, b_k + \epsilon 2^{-k})$ cover the closed interval $[a + \epsilon, b]$, and it follows by the Heine–Borel theorem [A13] that $[a + \epsilon, b] \subset \bigcup_{k=1}^{n}(a_k, b_k + \epsilon 2^{-k})$ for some n. But then $(a + \epsilon, b] \subset \bigcup_{k=1}^{n}(a_k, b_k + \epsilon 2^{-k}]$, and by the finite case, $b - (a + \epsilon) \leq \sum_{k=1}^{n}(b_k + \epsilon 2^{-k} - a_k) \leq \sum_{k=1}^{\infty}(b_k - a_k) + \epsilon$. Since ϵ was arbitrary, the result follows.   ∎

Theorem 1.3 will be the starting point for the theory of Lebesgue measure as developed in Sections 2 and 3. Taken together, parts (i) and (ii) of the theorem for only finitely many intervals I_k imply (1.4) for disjoint A and B. Like (1.4), they follow immediately from the additivity of the Riemann integral; but the point is to give an independent development of which the Riemann theory will be an eventual by-product.

To pass from the finite to the infinite case in part (i) of the theorem is easy. But to pass from the finite to the infinite case in part (ii) involves compactness, a profound idea underlying all of modern analysis. And it is part (ii) that shows that an interval I of positive length is not negligible: $|I|$ is

a positive lower bound for the sum of the lengths of the intervals in any covering of I.

The Measure Theory of Diophantine Approximation*

Diophantine approximation has to do with the approximation of real numbers x by rational fractions p/q. The measure theory of Diophantine approximation has to do with the degree of approximation that is possible if one disregards negligible sets of real x.

For each positive integer q, x must lie between some pair of successive multiples of $1/q$, so that for some p, $|x - p/q| \leq 1/q$. Since for each q the intervals

$$(1.32) \qquad \left(\frac{p}{q} - \frac{1}{2q}, \frac{p}{q} + \frac{1}{2q} \right] \qquad \cdot$$

decompose the line, the error of approximation can be further reduced to $1/2q$: For each q there is a p such that $|x - p/q| \leq 1/2q$. These observations are of course trivial. But for "most" real numbers x there will be many values of p and q for which x lies very near the center of the interval (1.32), so that p/q is a very sharp approximation to x.

Theorem 1.4. *If x is irrational, there are infinitely many irreducible fractions p/q such that*

$$(1.33) \qquad \left| x - \frac{p}{q} \right| < \frac{1}{q^2}.$$

This famous theorem of Dirichlet says that for infinitely many p and q, x lies in $(p/q - 1/q^2, p/q + 1/q^2)$ and hence is indeed very near the center of (1.32).

PROOF. For a positive integer Q, decompose $[0,1)$ into the Q subintervals $[(i-1)/Q, i/Q)$, $i = 1, \ldots, Q$. The points (fractional parts) $\{qx\} = qx - \lfloor qx \rfloor$ for $q = 0, 1, \ldots, Q$ lie in $[0,1)$, and since there are $Q + 1$ points[†] and only Q subintervals, it follows (Dirichlet's drawer principle) that some subinterval contains more than one point. Suppose that $\{q'x\}$ and $\{q''x\}$ lie in the same subinterval and $0 \leq q' < q'' \leq Q$. Take $q = q'' - q'$ and $p = \lfloor q''x \rfloor - \lfloor q'x \rfloor$; then $1 \leq q \leq Q$ and $|qx - p| = |\{q''x\} - \{q'x\}| < 1/Q$:

$$(1.34) \qquad \left| x - \frac{p}{q} \right| < \frac{1}{qQ} \leq \frac{1}{q^2}.$$

If p and q have any common factors, cancel them; this will not change the left side of (1.34), and it will decrease q.

For each Q, therefore, there is an irreducible p/q satisfying (1.34).[‡] Suppose there are only finitely many irreducible solutions of (1.33), say $p_1/q_1, \ldots, p_m/q_m$. Since x is irrational, the $|x - p_k/q_k|$ are all positive, and it is possible to choose Q so that Q^{-1} is smaller than each of them. But then the p/q of (1.34) is a solution of (1.33), and since $|x - p/q| < 1/Q$, there is a contradiction. ∎

---

*This topic may be omitted.

[†]Although the fact is not technically necessary to the proof, these points are distinct: $\{q'x\} = \{q''x\}$ implies $(q'' - q')x = \lfloor q''x \rfloor - \lfloor q'x \rfloor$, which in turn implies that x is rational unless $q' = q''$.

[‡]This much of the proof goes through even if x is rational.

In the measure theory of Diophantine approximation, one looks at the set of real x having such and such approximation properties and tries to show that this set is negligible or else that its complement is. Since the set of rationals is negligible, Theorem 1.4 implies such a result: Apart from a negligible set of x, (1.33) has infinitely many irreducible solutions.

What happens if the inequality (1.33) is tightened? Consider

$$(1.35) \qquad \left| x - \frac{p}{q} \right| < \frac{1}{q^2 \varphi(q)},$$

and let A_φ consist of the real x for which (1.35) has infinitely many irreducible solutions. Under what conditions on φ will A_φ have negligible complement? If $\varphi(q) \le 1$, then (1.35) is weaker than (1.33): $\varphi(q) > 1$ in the interesting cases. Since x satisfies (1.35) for infinitely many irreducible p/q if and only if $x - \lfloor x \rfloor$ does, A_φ may as well be redefined as the set of x in $(0,1)$ (or even as the set of irrational x in $(0,1)$) for which (1.35) has infinitely many solutions.

Theorem 1.5. *Suppose that φ is positive and nondecreasing. If*

$$(1.36) \qquad \sum_q \frac{1}{q\varphi(q)} = \infty,$$

then A_φ has negligible complement.

Theorem 1.4 covers the case $\varphi(q) \equiv 1$. Although this is the natural place to state Theorem 1.5 in its general form, the proof, which involves continued fractions and the ergodic theorem, must be postponed; see Section 24, p. 324. The converse, on the other hand, has a very simple proof.

Theorem 1.6. *Suppose that φ is positive. If*

$$(1.37) \qquad \sum_q \frac{1}{q\varphi(q)} < \infty,$$

then A_φ is negligible.

PROOF.   Given ϵ, choose q_0 so that $\sum_{q \ge q_0} 1/q\varphi(q) < \epsilon/4$. If $x \in A_\varphi$, then (1.35) holds for some $q \ge q_0$, and since $0 < x < 1$, the corresponding p lies in the range $0 \le p \le q$. Therefore,

$$A_\varphi \subset \bigcup_{q \ge q_0} \bigcup_{p=0}^{q} \left(\frac{p}{q} - \frac{1}{q^2\varphi(q)}, \frac{p}{q} + \frac{1}{q^2\varphi(q)} \right].$$

The right side here is a countable union of intervals covering A_φ, and the sum of their lengths is

$$\sum_{q \ge q_0} \sum_{p=0}^{q} \frac{2}{q^2\varphi(q)} = \sum_{q \ge q_0} \frac{2(q+1)}{q^2\varphi(q)} \le \sum_{q \ge q_0} \frac{4}{q\varphi(q)} < \epsilon.$$

Thus A_φ satisfies the definition ((1.22) and (1.23)) of negligibility.   ∎

If $\varphi_1(q) \equiv 1$, then (1.36) holds and hence A_{φ_1} has negligible complement (as follows also from Theorem 1.4). If $\varphi_2(q) = q^\epsilon$, however, then (1.37) holds and A_{φ_2} itself is negligible Outside the negligible set $A_{\varphi_1}^c \cup A_{\varphi_2}$, therefore, $|x - p/q| < 1/q^2$ has infinitely many irreducible solutions but $|x - p/q| < 1/q^{2+\epsilon}$ has only finitely many. Similarly, since $\Sigma_q 1/(q \log q)$ diverges but $\Sigma_q 1/(q \log^{1+\epsilon} q)$ converges, outside a negligible set $|x - p/q| < 1/(q^2 \log q)$ has infinitely many irreducible solutions but $|x - p/q| < 1/(q^2 \log^{1+\epsilon} q)$ has only finitely many.

Rational approximations to x obtained by truncating its binary (or decimal) expansion are very inaccurate: see Example 4.17. The sharp rational approximations to x come from truncation of its continued-fraction expansion: see Section 24.

PROBLEMS

Some problems involve concepts not required for an understanding of the text, or concepts treated only in later sections; there are no problems whose solutions are used in the text itself. An arrow ↑ points back to a problem (the one immediately preceding if no number is given) the solution and terminology of which are assumed. See Notes on the Problems, p. 552.

1.1. (a) Show that a *discrete* probability space (see Example 2.8 for the formal definition) cannot contain an infinite sequence $A_1, A_2, \ldots$ of independent events each of probability $\frac{1}{2}$. Since A_n could be identified with heads on the nth toss of a coin, the existence of such a sequence would make this section superfluous.

(b) Suppose that $0 \leq p_n \leq 1$, and put $\alpha_n = \min\{p_n, 1 - p_n\}$. Show that, if $\Sigma_n \alpha_n$ diverges, then no discrete probability space can contain independent events $A_1, A_2, \ldots$ such that A_n has probability p_n.

1.2. Show that N and N^c are dense [A15] in $(0, 1]$.

1.3. ↑ Define a set A to be *trifling*[†] if for each ϵ there exists a *finite* sequence of intervals I_k satisfying (1.22) and (1.23). This definition and the definition of negligibility apply as they stand to all sets on the real line, not just to subsets of $(0, 1]$.

(a) Show that a trifling set is negligible.

(b) Show that the closure of a trifling set is also trifling.

(c) Find a bounded negligible set that is not trifling.

(d) Show that the closure of a negligible set may not be negligible.

(e) Show that finite unions of trifling sets are trifling but that this can fail for countable unions.

1.4. ↑ For $i = 0, \ldots, r - 1$, let $A_r(i)$ be the set of numbers in $(0, 1]$ whose nonterminating expansions in the base r do not contain the digit i.

(a) Show that $A_r(i)$ is trifling.

(b) Find a trifling set A such that every point in the unit interval can be represented in the form $x + y$ with x and y in A.

[†]Like *negligible*, *trifling* is a nonce word used only here. The trifling sets are exactly the sets of content 0: See Problem 3.15

(c) Let $A_r(i_1, \ldots, i_k)$ consist of the numbers in the unit interval in whose base-r expansions the digits $i_1, \ldots, i_k$ nowhere appear consecutively in that order. Show that it is trifling. What does this imply about the monkey that types at random?

1.5. ↑ The *Cantor set* C can be defined as the closure of $A_3(1)$

(a) Show that C is uncountable but trifling.

(b) From $[0,1]$ remove the open middle third $(\frac{1}{3}, \frac{2}{3})$; from the remainder, a union of two closed intervals, remove the two open middle thirds $(\frac{1}{9}, \frac{2}{9})$ and $(\frac{7}{9}, \frac{8}{9})$. Show that C is what remains when this process is continued ad infinitum.

(c) Show that C is perfect [A15]

1.6. Put $M(t) = \int_0^1 e^{t s_n(\omega)} d\omega$, and show by successive differentiations under the integral that

(1.38)
$$M^{(k)}(0) = \int_0^1 s_n^k(\omega) \, d\omega.$$

Over each dyadic interval of rank n, $s_n(\omega)$ has a constant value of the form $\pm 1 \pm 1 \pm \cdots \pm 1$, and therefore $M(t) = 2^{-n} \sum \exp t(\pm 1 \pm 1 \pm \cdots \pm 1)$, where the sum extends over all 2^n n-long sequences of $+1$'s and -1's. Thus

(1.39)
$$M(t) = \left(\frac{e^t + e^{-t}}{2} \right)^n = (\cosh t)^n$$

Use this and (1.38) to give new proofs of (1.16), (1.18), and (1.28). (This, the method of moment generating functions, will be investigated systematically in Section 9.)

1.7. ↑ By an argument similar to that leading to (1.39) show that the Rademacher functions satisfy

$$\int_0^1 \exp\left[i \sum_{k=1}^n a_k r_k(\omega) \right] d\omega = \prod_{k=1}^n \frac{e^{i a_k} + e^{-i a_k}}{2}$$

$$= \prod_{k=1}^n \cos a_k.$$

Take $a_k = t 2^{-k}$, and from $\sum_{k=1}^\infty r_k(\omega) 2^{-k} = 2\omega - 1$ deduce

(1.40)
$$\frac{\sin t}{t} = \prod_{k=1}^\infty \cos \frac{t}{2^k}$$

by letting $n \to \infty$ inside the integral above. Derive Vieta's formula

$$\frac{2}{\pi} = \frac{\sqrt{2}}{2} \frac{\sqrt{2 + \sqrt{2}}}{2} \frac{\sqrt{2 + \sqrt{2 + \sqrt{2}}}}{2} \cdots$$

1.8. A number ω is normal in the base 2 if and only if for each positive ϵ there exists an $n_0(\epsilon, \omega)$ such that $|n^{-1} \sum_{i=1}^n d_i(\omega) - \frac{1}{2}| < \epsilon$ for all n exceeding $n_0(\epsilon, \omega)$.

Theorem 1.2 concerns the entire dyadic expansion, whereas Theorem 1.1 concerns only the beginning segment. Point up the difference by showing that for $\epsilon < \frac{1}{2}$ the $n_0(\epsilon, \omega)$ above cannot be the same for all ω in N—in other words, $n^{-1}\sum_{i=1}^{n}d_i(\omega)$ converges to $\frac{1}{2}$ for all ω in N, but not uniformly. But see Problem 13.9.

1.9. 1.3↑ (a) Using the finite form of Theorem 1.3(ii), together with Problem 1.3(b), show that a trifling set is nowhere dense [A15].

(b) Put $B = \bigcup_n (r_n - 2^{-n-2}, r_n + 2^{-n-2}]$, where $r_1, r_2, \ldots$ is an enumeration of the rationals in $(0, 1]$. Show that $(0, 1] - B$ is nowhere dense but not trifling or even negligible.

(c) Show that a compact negligible set is trifling

1.10. ↑ A set of the first category [A15] can be represented as a countable union of nowhere dense sets; this is a topological notion of smallness, just as negligibility is a metric notion of smallness. Neither condition implies the other:

(a) Show that the nonnegligible set N of normal numbers is of the first category by proving that $A_m = \bigcap_{n=m}^{\infty}[\omega \colon |n^{-1}s_n(\omega)| < \frac{1}{2}]$ is nowhere dense and $N \subset \bigcup_m A_m$.

(b) According to a famous theorem of Baire, a nonempty interval is *not* of the first category. Use this fact to prove that the negligible set $N^c = (0, 1] - N$ is not of the first category.

1.11. Prove:

(a) If x is rational, (1.33) has only finitely many irreducible solutions

(b) Suppose that $\varphi(q) \geq 1$ and (1.35) holds for infinitely many pairs p, q but only for finitely many relatively prime ones. Then x is rational.

(c) If φ goes to infinity too rapidly, then A_φ is negligible (Theorem 1.6). But however rapidly φ goes to infinity, A_φ is nonempty, even uncountable. *Hint*· Consider $x = \sum_{k=1}^{\infty} 1/2^{\alpha(k)}$ for integral $\alpha(k)$ increasing very rapidly to infinity.

SECTION 2. PROBABILITY MEASURES

Spaces

Let Ω be an arbitrary space or set of points ω. In probability theory Ω consists of all the possible results or outcomes ω of an experiment or observation. For observing the number of heads in n tosses of a coin the space Ω is $\{0, 1, \ldots, n\}$; for describing the complete history of the n tosses Ω is the space of all 2^n n-long sequences of H's and T's; for an infinite sequence of tosses Ω can be taken as the unit interval as in the preceding section; for the number of α-particles emitted by a substance during a unit interval of time or for the number of telephone calls arriving at an exchange Ω is $\{0, 1, 2, \ldots\}$; for the position of a particle Ω is three-dimensional Euclidean space; for describing the motion of the particle Ω is an appropriate space of functions; and so on. Most Ω's to be considered are interesting from the point of view of geometry and analysis as well as that of probability.

Viewed probabilistically, a subset of Ω is an *event* and an element ω of Ω is a *sample point*.

Assigning Probabilities

In setting up a space Ω as a probabilistic model, it is natural to try and assign probabilities to as many events as possible. Consider again the case $\Omega = (0, 1]$ —the unit interval. It is natural to try and go beyond the definition (1.3) and assign probabilities in a systematic way to sets other than finite unions of intervals. Since the set of nonnormal numbers is negligible, for example, one feels it ought to have probability 0. For another probabilistically interesting set that is not a finite union of intervals, consider

$$(2.1) \qquad \bigcup_{n=1}^{\infty} \left[\omega\colon -a < s_1(\omega), \ldots, s_{n-1}(\omega) < b, \, s_n(\omega) = -a \right],$$

where a and b are positive integers. This is the event that the gambler's fortune reaches $-a$ before it reaches $+b$; it represents ruin for a gambler with a dollars playing against an adversary with b dollars, the rule being that they play until one or the other runs out of capital.

The union in (2.1) is countable and disjoint, and for each n the set in the union is itself a union of certain of the intervals (1.9). Thus (2.1) is a countably infinite disjoint union of intervals, and it is natural to take as its probability the sum of the lengths of these constituent intervals. Since the set of normal numbers is not a countable disjoint union of intervals, however, this extension of the definition of probability would still not cover all the interesting sets (events) in $(0, 1]$.

It is, in fact, not fruitful to try to predict just which sets probabilistic analysis will require and then assign probabilities to them in some *ad hoc* way. The successful procedure is to develop a general theory that assigns probabilities at once to the sets of a class so extensive that most of its members never actually arise in probability theory. That being so, why not ask for a theory that goes all the way and applies to *every* set in a space Ω? In the case of the unit interval, should there not exist a well-defined probability that the random point ω lies in A, whatever the set A may be? The answer turns out to be no (see p. 45), and it is necessary to work within subclasses of the class of all subsets of a space Ω. The classes of the appropriate kinds—the fields and σ-fields—are defined and studied in this section. The theory developed here covers the spaces listed above, including the unit interval, and a great variety of others.

Classes of Sets

It is necessary to single out for special treatment classes of subsets of a space Ω, and to be useful, such a class must be closed under various of the

operations of set theory. Once again the unit interval provides an instructive example.

Example 2.1.* Consider the set N of normal numbers in the form (1.24), where $s_n(\omega)$ is the sum of the first n Rademacher functions. Since a point ω lies in N if and only if $\lim_n n^{-1}s_n(\omega) = 0$, N can be put in the form

$$(2\,2) \qquad N = \bigcap_{k=1}^{\infty} \bigcup_{m=1}^{\infty} \bigcap_{n=m}^{\infty} \left[\omega: \left| n^{-1}s_n(\omega) \right| < k^{-1} \right].$$

Indeed, because of the very meaning of union and of intersection, ω lies in the set on the right here if and only if for every k there exists an m such that $|n^{-1}s_n(\omega)| < k^{-1}$ holds for all $n \geq m$, and this is just the definition of convergence to 0—with the usual ϵ replaced by k^{-1} to avoid the formation of an uncountable intersection. Since $s_n(\omega)$ is constant over each dyadic interval of rank n, the set $[\omega: n^{-1}s_n(\omega)| < k^{-1}]$ is a finite disjoint union of intervals. The formula (2.2) shows explicitly how N is constructed in steps from these simpler sets. ■

A systematic treatment of the ideas in Section 1 thus requires a class of sets that contains the intervals and is closed under the formation of countable unions and intersections. Note that a singleton [A1] $\{x\}$ is a countable intersection $\bigcap_n (x - n^{-1}, x]$ of intervals. If a class contains all the singletons and is closed under the formation of *arbitrary* unions, then of course it contains *all* the subsets of Ω. As the theory of this section and the next does not apply to such extensive classes of sets, attention must be restricted to countable set-theoretic operations and in some cases even to finite ones.

Consider now a completely arbitrary nonempty space Ω. A class $\mathscr{F}$ of subsets of Ω is called a *field*† if it contains Ω itself and is closed under the formation of complements and finite unions:

(i) $\Omega \in \mathscr{F}$;

(ii) $A \in \mathscr{F}$ implies $A^c \in \mathscr{F}$;

(iii) $A, B \in \mathscr{F}$ implies $A \cup B \in \mathscr{F}$.

Since Ω and the empty set $\varnothing$ are complementary, (i) is the same in the presence of (ii) as the assumption $\varnothing \in \mathscr{F}$. In fact, (i) simply ensures that $\mathscr{F}$ is nonempty: If $A \in \mathscr{F}$, then $A^c \in \mathscr{F}$ by (ii) and $\Omega = A \cup A^c \in \mathscr{F}$ by (iii).

By DeMorgan's law, $A \cap B = (A^c \cup B^c)^c$ and $A \cup B = (A^c \cap B^c)^c$. If $\mathscr{F}$ is closed under complementation, therefore, it is closed under the formation of finite unions if and only if it is closed under the formation of finite intersec-

---

*Many of the examples in the book simply illustrate the concepts at hand, but others contain definitions and facts needed subsequently

†The term *algebra* is often used in place of *field*

tions. Thus (iii) can be replaced by the requirement

(iii′) $A, B \in \mathscr{F}$ implies $A \cap B \in \mathscr{F}$.

A class $\mathscr{F}$ of subsets of Ω is a *σ-field* if it is a field and if it is also closed under the formation of *countable* unions:

(iv) $A_1, A_2, \ldots \in \mathscr{F}$ implies $A_1 \cup A_2 \cup \cdots \in \mathscr{F}$.

By the infinite form of DeMorgan's law, assuming (iv) is the same thing as assuming

(iv′) $A_1, A_2, \ldots \in \mathscr{F}$ implies $A_1 \cap A_2 \cap \cdots \in \mathscr{F}$.

Note that (iv) implies (iii) because one can take $A_1 = A$ and $A_n = B$ for $n \geq 2$. A field is sometimes called a *finitely additive* field to stress that it need not be a σ-field. A set in a given class $\mathscr{F}$ is said to be *measurable* $\mathscr{F}$ or to be an $\mathscr{F}$-set. A field or σ-field of subsets of Ω will sometimes be called a field or σ-field *in* Ω.

Example 2.2. Section 1 began with a consideration of the sets (1.2), the finite disjoint unions of subintervals of $\Omega = (0, 1]$. Augmented by the empty set, this class is a field $\mathscr{B}_0$: Suppose that $A = (a_1, a'_1] \cup \cdots \cup (a_m, a'_m]$, where the notation is so chosen that $a_1 \leq \cdots \leq a_m$. If the $(a_i, a'_i]$ are disjoint, then A^c is $(0, a_1] \cup (a'_1, a_2] \cup \cdots \cup (a'_{m-1}, a_m] \cup (a'_m, 1]$ and so lies in $\mathscr{B}_0$ (some of these intervals may be empty, as a'_i and a_{i+1} may coincide). If $B = (b_1, b'_1] \cup \cdots \cup (b_n, b'_n]$, the $(b_j, b'_j]$ again disjoint, then $A \cap B = \bigcup_{i=1}^{m} \bigcup_{j=1}^{n} \{(a_i, a'_i] \cap (b_j, b'_j]\}$; each intersection here is again an interval or else the empty set, and the union is disjoint, and hence $A \cap B$ is in $\mathscr{B}_0$. Thus $\mathscr{B}_0$ satisfies (i), (ii), and (iii′).

Although $\mathscr{B}_0$ is a field, it is not a σ-field: It does not contain the singletons $\{x\}$, even though each is a countable intersection $\bigcap_n (x - n^{-1}, x]$ of $\mathscr{B}_0$-sets. And $\mathscr{B}_0$ does not contain the set (2.1), a countable union of intervals that cannot be represented as a finite union of intervals. The set (2.2) of normal numbers is also outside $\mathscr{B}_0$.   ∎

The definitions above involve distinctions perhaps most easily made clear by a pair of artificial examples.

Example 2.3. Let $\mathscr{F}$ consist of the finite and the cofinite sets (A being cofinite if A^c is finite). Then $\mathscr{F}$ is a field. If Ω is finite, then $\mathscr{F}$ contains all the subsets of Ω and hence is a σ-field as well. If Ω is infinite, however, then $\mathscr{F}$ is not a σ-field. Indeed, choose in Ω a set A that is countably infinite and has infinite complement. (For example, choose a sequence $\omega_1, \omega_2, \ldots$ of distinct points in Ω and take $A = \{\omega_2, \omega_4, \ldots\}$.) Then $A \notin \mathscr{F}$, even though

A is the union, necessarily countable, of the singletons it contains and each singleton is in $\mathscr{F}$. This shows that the definition of σ-field is indeed more restrictive than that of field. ∎

Example 2.4. Let $\mathscr{F}$ consist of the countable and the cocountable sets (A being cocountable if A^c is countable). Then $\mathscr{F}$ is a σ-field. If Ω is uncountable, then it contains a set A such that A and A^c are both uncountable.[†] Such a set is not in $\mathscr{F}$, which shows that even a σ-field may not contain all the subsets of Ω; furthermore, this set is the union (uncountable) of the singletons it contains and each singleton is in $\mathscr{F}$, which shows that a σ-field may not be closed under the formation of arbitrary unions. ∎

The largest σ-field in Ω is the *power class* 2^{Ω}, consisting of *all* the subsets of Ω; the smallest σ-field consists only of the empty set and Ω itself.

The elementary facts about fields and σ-fields are easy to prove: If $\mathscr{F}$ is a field, then $A, B \in \mathscr{F}$ implies $A - B = A \cap B^c \in \mathscr{F}$ and $A \triangle B = (A - B) \cup (B - A) \in \mathscr{F}$. Further, it follows by induction on n that $A_1, \ldots, A_n \in \mathscr{F}$ implies $A_1 \cup \cdots \cup A_n \in \mathscr{F}$ and $A_1 \cap \cdots \cap A_n \in \mathscr{F}$.

A field is closed under the finite set-theoretic operations, and a σ-field is closed also under the countable ones. The analysis of a probability problem usually begins with the sets of some rather small class $\mathscr{A}$, such as the class of subintervals of $(0, 1]$. As in Example 2.1, probabilistically natural constructions involving finite and countable operations can then lead to sets outside the initial class $\mathscr{A}$. This leads one to consider a class of sets that (i) contains $\mathscr{A}$ and (ii) is a σ-field; it is natural and convenient, as it turns out, to consider a class that has these two properties and that in addition (iii) is in a certain sense as small as possible. As will be shown, this class is the *intersection of all the σ-fields containing $\mathscr{A}$*; it is called the *σ-field generated by* $\mathscr{A}$ and is denoted by $\sigma(\mathscr{A})$.

There do exist σ-fields containing $\mathscr{A}$, the class of all subsets of Ω being one. Moreover, a completely arbitrary intersection of σ-fields (however many of them there may be) is itself a σ-field: Suppose that $\mathscr{F} = \bigcap_{\theta} \mathscr{F}_{\theta}$, where θ ranges over an arbitrary index set and each $\mathscr{F}_{\theta}$ is a σ-field. Then $\Omega \in \mathscr{F}_{\theta}$ for all θ, so that $\Omega \in \mathscr{F}$. And $A \in \mathscr{F}$ implies for each θ that $A \in \mathscr{F}_{\theta}$ and hence $A^c \in \mathscr{F}_{\theta}$, so that $A^c \in \mathscr{F}$. If $A_n \in \mathscr{F}$ for each n, then $A_n \in \mathscr{F}_{\theta}$ for each n and θ, so that $\bigcup_n A_n$ lies in each $\mathscr{F}_{\theta}$ and hence in $\mathscr{F}$.

Thus the intersection in the definition of $\sigma(\mathscr{A})$ is indeed a σ-field containing $\mathscr{A}$. It is as small as possible, in the sense that it is contained in every σ-field that contains $\mathscr{A}$: if $\mathscr{A} \subset \mathscr{G}$ and $\mathscr{G}$ is a σ-field, then $\mathscr{G}$ is one of

---

[†] If Ω is the unit interval, for example, take $A = (0, \frac{1}{2}]$, say. To show that the general uncountable Ω contains such an A requires the axiom of choice [A8]. As a matter of fact, to prove the existence of the sequence alluded to in Example 2.3 requires a form of the axiom of choice, as does even something so apparently down-to-earth as proving that a countable union of negligible sets is negligible. Most of us use the axiom of choice completely unaware of the fact Even Borel and Lebesgue did; see WAGON, pp. 217 ff.

the σ-fields in the intersection defining $\sigma(\mathscr{A})$, so that $\sigma(\mathscr{A}) \subset \mathscr{G}$. Thus $\sigma(\mathscr{A})$ has these three properties:

(i) $\mathscr{A} \subset \sigma(\mathscr{A})$;

(ii) $\sigma(\mathscr{A})$ *is a σ-field*;

(iii) *if $\mathscr{A} \subset \mathscr{G}$ and $\mathscr{G}$ is a σ-field, then $\sigma(\mathscr{A}) \subset \mathscr{G}$.*

The importance of σ-fields will gradually become clear.

Example 2.5. If $\mathscr{F}$ is a σ-field, then obviously $\sigma(\mathscr{F}) = \mathscr{F}$. If $\mathscr{A}$ consists of the singletons, then $\sigma(\mathscr{A})$ is the σ-field in Example 2.4. If $\mathscr{A}$ is empty or $\mathscr{A} = \{\varnothing\}$ or $\mathscr{A} = \{\Omega\}$, then $\sigma(\mathscr{A}) = \{\varnothing, \Omega\}$. If $\mathscr{A} \subset \mathscr{A}'$, then $\sigma(\mathscr{A}) \subset \sigma(\mathscr{A}')$. If $\mathscr{A} \subset \mathscr{A}' \subset \sigma(\mathscr{A})$, then $\sigma(\mathscr{A}) = \sigma(\mathscr{A}')$.   ◼

Example 2.6. Let $\mathscr{I}$ be the class of subintervals of $\Omega = (0, 1]$, and define $\mathscr{B} = \sigma(\mathscr{I})$. The elements of $\mathscr{B}$ are called the *Borel sets* of the unit interval. The field $\mathscr{B}_0$ of Example 2.2 satisfies $\mathscr{I} \subset \mathscr{B}_0 \subset \mathscr{B}$, and hence $\sigma(\mathscr{B}_0) = \mathscr{B}$.

Since $\mathscr{B}$ contains the intervals and is a σ-field, repeated finite and countable set-theoretic operations starting from intervals will never lead outside $\mathscr{B}$. Thus $\mathscr{B}$ contains the set (2.2) of normal numbers. It also contains for example the open sets in $(0, 1]$: If G is open and $x \in G$, then there exist rationals a_x and b_x such that $x \in (a_x, b_x] \subset G$. But then $G = \bigcup_{x \in G} (a_x, b_x]$; since there are only countably many intervals with rational endpoints, G is a *countable* union of elements of $\mathscr{I}$ and hence lies in $\mathscr{B}$.

In fact, $\mathscr{B}$ contains all the subsets of $(0, 1]$ actually encountered in ordinary analysis and probability. It is large enough for all "practical" purposes. It does not contain every subset of the unit interval, however; see the end of Section 3 (p. 45). The class $\mathscr{B}$ will play a fundamental role in all that follows.   ◼

Probability Measures

A *set function* is a real-valued function defined on some class of subsets of Ω. A set function P on a field $\mathscr{F}$ is a *probability measure* if it satisfies these conditions:

(i) $0 \le P(A) \le 1$ for $A \in \mathscr{F}$;

(ii) $P(\varnothing) = 0$, $P(\Omega) = 1$;

(iii) if $A_1, A_2, \ldots$ is a disjoint sequence of $\mathscr{F}$-sets and if $\bigcup_{k=1}^{\infty} A_k \in \mathscr{F}$, then[†]

$$(2.3) \qquad P\left(\bigcup_{k=1}^{\infty} A_k \right) = \sum_{k=1}^{\infty} P(A_k).$$

[†]As the left side of (2 3) is invariant under permutations of the A_n, the same must be true of the right side. But in fact, according to Dirichlet's theorem [A26], a nonnegative series has the same value whatever order the terms are summed in.

The condition imposed on the set function P by (iii) is called *countable additivity*. Note that, since $\mathscr{F}$ is a field but perhaps not a σ-field, it is necessary in (iii) to assume that $\bigcup_{k=1}^{\infty} A_k$ lies in $\mathscr{F}$. If $A_1, \ldots, A_n$ are disjoint $\mathscr{F}$-sets, then $\bigcup_{k=1}^{n} A_k$ is also in $\mathscr{F}$ and (2.3) with $A_{n+1} = A_{n+2} = \cdots = \varnothing$ gives

$$(2.4) \qquad P\left(\bigcup_{k=1}^{n} A_k \right) = \sum_{k=1}^{n} P(A_k).$$

The condition that (2.4) holds for disjoint $\mathscr{F}$-sets is *finite additivity*; it is a consequence of countable additivity. It follows by induction on n that P is finitely additive if (2.4) holds for $n = 2$—if $P(A \cup B) = P(A) + P(B)$ for disjoint $\mathscr{F}$-sets A and B.

The conditions above are redundant, because (i) can be replaced by $P(A) \geq 0$ and (ii) by $P(\Omega) = 1$. Indeed, the weakened forms (together with (iii)) imply that $P(\Omega) = P(\Omega) + P(\varnothing) + P(\varnothing) + \cdots$, so that $P(\varnothing) = 0$, and $1 = P(\Omega) = P(A) + P(A^c)$, so that $P(A) \leq 1$.

Example 2.7. Consider as in Example 2.2 the field $\mathscr{B}_0$ of finite disjoint unions of subintervals of $\Omega = (0, 1]$. The definition (1.3) assigns to each $\mathscr{B}_0$-set a number—the sum of the lengths of the constituent intervals—and hence specifies a set function P on $\mathscr{B}_0$. Extended inductively, (1.4) says that P is finitely additive. In Section 1 this property was deduced from the additivity of the Riemann integral (see (1.5)). In Theorem 2.2 below, the finite additivity of P will be proved from first principles, and it will be shown that P is, in fact, countably additive—is a probability measure on the field $\mathscr{B}_0$. The hard part of the argument is in the proof of Theorem 1.3, already done; the rest will be easy. ∎

If $\mathscr{F}$ is a σ-field in Ω and P is a probability measure on $\mathscr{F}$, the triple $(\Omega, \mathscr{F}, P)$ is called a *probability measure space*, or simply a *probability space*. A *support* of P is any $\mathscr{F}$-set A for which $P(A) = 1$.

Example 2.8. Let $\mathscr{F}$ be the σ-field of all subsets of a countable space Ω, and let $p(\omega)$ be a nonnegative function on Ω. Suppose that $\sum_{\omega \in \Omega} p(\omega) = 1$, and define $P(A) = \sum_{\omega \in A} p(\omega)$; since $p(\omega) \geq 0$, the order of summation is irrelevant by Dirichlet's theorem [A26]. Suppose that $A = \bigcup_{i=1}^{\infty} A_i$, where the A_i are disjoint, and let $\omega_{i1}, \omega_{i2}, \ldots$ be the points in A_i. By the theorem on nonnegative double series [A27], $P(A) = \sum_{ij} p(\omega_{ij}) = \sum_i \sum_j p(\omega_{ij}) = \sum_i P(A_i)$, and so P is countably additive. This $(\Omega, \mathscr{F}, P)$ is a *discrete probability space*. It is the formal basis for discrete probability theory. ∎

Example 2.9. Now consider a probability measure P on an arbitrary σ-field $\mathscr{F}$ in an arbitrary space Ω; P is a *discrete probability measure* if there exist finitely or countably many points ω_k and masses m_k such that $P(A) = \sum_{\omega_k \in A} m_k$ for A in $\mathscr{F}$. Here P is discrete, but the space itself may not be. In

terms of indicator functions, the defining condition is $P(A) = \sum_k m_k I_A(\omega_k)$ for $A \in \mathcal{F}$. If the set $\{\omega_1, \omega_2, \ldots\}$ lies in $\mathcal{F}$, then it is a support of P.

If there is just one of these points, say ω_0, with mass $m_0 = 1$, then P is a *unit mass* at ω_0. In this case $P(A) = I_A(\omega_0)$ for $A \in \mathcal{F}$. ∎

Suppose that P is a probability measure on a field $\mathcal{F}$, and that $A, B \in \mathcal{F}$ and $A \subset B$. Since $P(A) + P(B - A) = P(B)$, P is *monotone*:

$$(2.5) \qquad\qquad P(A) \leq P(B) \qquad \text{if } A \subset B.$$

It follows further that $P(B - A) = P(B) - P(A)$, and as a special case,

$$(2.6) \qquad\qquad P(A^c) = 1 - P(A).$$

Other formulas familiar from the discrete theory are easily proved. For example,

$$(2.7) \qquad\qquad P(A) + P(B) = P(A \cup B) + P(A \cap B),$$

the common value of the two sides being $P(A \cup B^c) + 2P(A \cap B) + P(A^c \cap B)$. Subtraction gives

$$(2.8) \qquad\qquad P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

This is the case $n = 2$ of the general *inclusion-exclusion formula*:

$$(2.9) \quad P\left(\bigcup_{k=1}^{n} A_k\right) = \sum_i P(A_i) - \sum_{i<j} P(A_i \cap A_j)$$

$$+ \sum_{i<j<k} P(A_i \cap A_j \cap A_k) + \cdots + (-1)^{n+1} P(A_1 \cap \cdots \cap A_n).$$

To deduce this inductively from (2.8), note that (2.8) gives

$$P\left(\bigcup_{k=1}^{n+1} A_k\right) = P\left(\bigcup_{k=1}^{n} A_k\right) + P(A_{n+1}) - P\left(\bigcup_{k=1}^{n} (A_k \cap A_{n+1})\right).$$

Applying (2.9) to the first and third terms on the right gives (2.9) with $n + 1$ in place of n.

If $B_1 = A_1$ and $B_k = A_k \cap A_1^c \cap \cdots \cap A_{k-1}^c$, then the B_k are disjoint and $\bigcup_{k=1}^{n} A_k = \bigcup_{k=1}^{n} B_k$, so that $P(\bigcup_{k=1}^{n} A_k) = \sum_{k=1}^{n} P(B_k)$. Since $P(B_k) \leq P(A_k)$ by monotonicity, this establishes the *finite subadditivity* of P:

$$(2.10) \qquad\qquad P\left(\bigcup_{k=1}^{n} A_k\right) \leq \sum_{k=1}^{n} P(A_k).$$

Here, of course, the A_k need not be disjoint. Sometimes (2.10) is called *Boole's inequality*.

In these formulas all the sets are naturally assumed to lie in the field $\mathcal{F}$. The derivations above involve only the finite additivity of P. Countable additivity gives further properties:

Theorem 2.1. *Let P be a probability measure on a field $\mathcal{F}$.*

(i) *Continuity from below*: If A_n and A lie in $\mathcal{F}$ and[†] $A_n \uparrow A$, then $P(A_n) \uparrow P(A)$.

(ii) *Continuity from above*: If A_n and A lie in $\mathcal{F}$ and $A_n \downarrow A$, then $P(A_n) \downarrow P(A)$.

(iii) *Countable subadditivity*: If $A_1, A_2, \ldots$ and $\bigcup_{k=1}^{\infty} A_k$ lie in $\mathcal{F}$ (the A_k need not be disjoint), then

$$(2.11) \qquad P\left(\bigcup_{k=1}^{\infty} A_k \right) \le \sum_{k=1}^{\infty} P(A_k).$$

PROOF. For (i), put $B_1 = A_1$ and $B_k = A_k - A_{k-1}$. Then the B_k are disjoint, $A = \bigcup_{k=1}^{\infty} B_k$, and $A_n = \bigcup_{k=1}^{n} B_k$, so that by countable and finite additivity, $P(A) = \sum_{k=1}^{\infty} P(B_k) = \lim_n \sum_{k=1}^{n} P(B_k) = \lim_n P(A_n)$. For (ii), observe that $A_n \downarrow A$ implies $A_n^c \uparrow A^c$, so that $1 - P(A_n) \uparrow 1 - P(A)$.

As for (iii), increase the right side of (2.10) to $\sum_{k=1}^{\infty} P(A_k)$ and then apply part (i) to the left side. ∎

Example 2.10. In the presence of finite additivity, a special case of (ii) implies countable additivity. *If P is a finitely additive probability measure on the field $\mathcal{F}$, and if $A_n \downarrow \varnothing$ for sets A_n in $\mathcal{F}$ implies $P(A_n) \downarrow 0$, then P is countably additive.* Indeed, if $B = \bigcup_k B_k$ for disjoint sets B_k (B and the B_k in $\mathcal{F}$), then $C_n = \bigcup_{k > n} B_k = B - \bigcup_{k \le n} B_k$ lies in the field $\mathcal{F}$, and $C_n \downarrow \varnothing$. The hypothesis, together with finite additivity, gives $P(B) - \sum_{k=1}^{n} P(B_k) = P(C_n) \to 0$, and hence $P(B) = \sum_{k=1}^{\infty} P(B_k)$. ∎

Lebesgue Measure on the Unit Interval

The definition (1.3) specifies a set function on the field $\mathcal{B}_0$ of finite disjoint unions of intervals in $(0, 1]$; the problem is to prove P countably additive. It will be convenient to change notation from P to λ, and to denote by $\mathcal{F}$ the class of subintervals $(a, b]$ of $(0, 1]$; then $\lambda(I) = |I| = b - a$ is ordinary length. Regard $\varnothing$ as an element of $\mathcal{F}$ of length 0. If $A = \bigcup_{i=1}^{n} I_i$, the I_i being

---

[†]For the notation, see [A4] and [A10].

disjoint $\mathscr{J}$-sets, the definition (1.3) in the new notation is

$$(2.12) \qquad \lambda(A) = \sum_{i=1}^{n} \lambda(I_i) = \sum_{i=1}^{n} |I_i|.$$

As pointed out in Section 1, there is a question of uniqueness here, because A will have other representations as a finite disjoint union $\bigcup_{j=1}^{m} J_j$ of $\mathscr{J}$-sets. But $\mathscr{J}$ is closed under the formation of finite intersections, and so the finite form of Theorem 1.3(iii) gives

$$(2.13) \qquad \sum_{i=1}^{n} |I_i| = \sum_{i=1}^{n} \sum_{j=1}^{m} |I_i \cap J_j| = \sum_{j=1}^{m} |J_j|.$$

(Some of the $I_i \cap J_j$ may be empty, but the corresponding lengths are then 0.) The definition is indeed consistent.

Thus (2.12) defines a set function λ on $\mathscr{B}_0$, a set function called *Lebesgue measure*.

Theorem 2.2. *Lebesgue measure λ is a (countably additive) probability measure on the field $\mathscr{B}_0$.*

PROOF. Suppose that $A = \bigcup_{k=1}^{\infty} A_k$, where A and the A_k are $\mathscr{B}_0$-sets and the A_k are disjoint. Then $A = \bigcup_{i=1}^{n} I_i$ and $A_k = \bigcup_{j=1}^{m_k} J_{kj}$ are disjoint unions of $\mathscr{J}$-sets, and (2.12) and Theorem 1.3(iii) give

$$(2.14) \qquad \lambda(A) = \sum_{i=1}^{n} |I_i| = \sum_{i=1}^{n} \sum_{k=1}^{\infty} \sum_{j=1}^{m_k} |I_i \cap J_{kj}|$$

$$= \sum_{k=1}^{\infty} \sum_{j=1}^{m_k} |J_{kj}| = \sum_{k=1}^{\infty} \lambda(A_k). \qquad \blacksquare$$

In Section 3 it is shown how to extend λ from $\mathscr{B}_0$ to the larger class $\mathscr{B} = \sigma(\mathscr{B}_0)$ of Borel sets in $(0, 1]$. This will complete the construction of λ as a probability measure (countably additive, that is) on $\mathscr{B}$. and the construction is fundamental to all that follows. For example, the set N of normal numbers lies in $\mathscr{B}$ (Example 2.6), and it will turn out that $\lambda(N) = 1$, as probabilistic intuition requires. (In Chapter 2, λ will be defined for sets outside the unit interval as well.)

It is well to pause here and consider just what is involved in the construction of Lebesgue measure on the Borel sets of the unit interval. That length defines a finitely additive set function on the class $\mathscr{J}$ of intervals in $(0, 1]$ is a consequence of Theorem 1.3 for the case of only finitely many intervals and thus involves only the most elementary properties of the real number system. But proving countable additivity on $\mathscr{J}$ requires the deeper property of

compactness (the Heine-Borel theorem). Once λ has been proved countably additive on $\mathcal{I}$, extending it to $\mathcal{B}_0$ by the definition (2.12) presents no real difficulty: the arguments involving (2.13) and (2.14) are easy. Difficulties again arise, however, in the further extension of λ from $\mathcal{B}_0$ to $\mathcal{B} = \sigma(\mathcal{B}_0)$, and here new ideas are again required. These ideas are the subject of Section 3, where it is shown that any probability measure on any field can be extended to the generated σ-field.

Sequence Space*

Let S be a finite set of points regarded as the possible outcomes of a simple observation or experiment. For tossing a coin, S can be $\{H, T\}$ or $\{0, 1\}$; for rolling a die, $S = \{1, \ldots, 6\}$; in information theory, S plays the role of a finite alphabet. Let $\Omega = S^\infty$ be the space of all infinite sequences

$$(2.15) \qquad \omega = (z_1(\omega), z_2(\omega), \ldots)$$

of elements of S: $z_k(\omega) \in S$ for all $\omega \in S^\infty$ and $k \geq 1$. The sequence (2.15) can be viewed as the result of repeating infinitely often the simple experiment represented by S. For $S = \{0, 1\}$, the space S^∞ is closely related to the unit interval; compare (1.8) and (2.15).

The space S^∞ is an infinite-dimensional Cartesian product. Each $z_k(\cdot)$ is a mapping of S^∞ onto S; these are the *coordinate functions*, or the *natural projections*. Let $S^n = S \times \cdots \times S$ be the Cartesian product of n copies of S; it consists of the n-long sequences $(u_1, \ldots, u_n)$ of elements of S. For such a sequence, the set

$$(2.16) \qquad \left[\omega: (z_1(\omega), \ldots, z_n(\omega)) = (u_1, \ldots, u_n) \right]$$

represents the event that the first n repetitions of the experiment give the outcomes $u_1, \ldots, u_n$ in sequence. A *cylinder of rank n* is a set of the form

$$(2.17) \qquad A = \left[\omega: (z_1(\omega), \ldots, z_n(\omega)) \in H \right],$$

where $H \subset S^n$. Note that A is nonempty if H is. If H is a singleton in S^n, (2.17) reduces to (2.16), which can be called a *thin* cylinder.

Let $\mathcal{C}_0$ be the class of cylinders of all ranks. Then $\mathcal{C}_0$ *is a field*: S^∞ and the empty set have the form (2.17) for $H = S^n$ and for $H = \varnothing$. If H is replaced by $S^n - H$, then (2.17) goes into its complement, and hence $\mathcal{C}_0$ is

---

*The ideas that follow are basic to probability theory and are used further on, in particular in Section 24 and (in more elaborate form) Section 36. On a first reading, however, one might prefer to skip to Section 3 and return to this topic as the need arises.

closed under complementation As for unions, consider (2.17) together with

$$(2.18) \qquad B = \left[\omega : \left(z_1(\omega), \dots, z_m(\omega) \right) \in I \right],$$

a cylinder of rank m. Suppose that $n \le m$ (symmetry); if H' consists of the sequences $(u_1, \dots, u_m)$ in S^m for which the truncated sequence $(u_1, \dots, u_n)$ lies in H, then (2.17) has the alternative form

$$(2.19) \qquad A = \left[\omega : \left(z_1(\omega), \dots, z_m(\omega) \right) \in H' \right].$$

Since it is now clear that

$$(2.20) \qquad A \cup B = \left[\omega : \left(z_i(\omega), \dots, z_m(\omega) \right) \in H' \cup I \right]$$

is also a cylinder, $\mathscr{C}_0$ is closed under the formation of finite unions and hence is indeed a field.

Let p_u, $u \in S$, be probabilities on S—nonnegative and summing to 1. Define a set function P on $\mathscr{C}_0$ (it will turn out to be a probability measure) in this way: For a cylinder A given by (2.17), take

$$(2.21) \qquad P(A) = \sum_H p_{u_1} \cdots p_{u_n},$$

the sum extending over all the sequences $(u_1, \dots, u_n)$ in H. As a special case,

$$(2.22) \qquad P\left[\omega : \left(z_1(\omega), \dots, z_n(\omega) \right) = (u_1, \dots, u_n) \right] = p_{u_1} \cdots p_{u_n}.$$

Because of the products on the right in (2.21) and (2.22), P is called *product measure*; it provides a model for an infinite sequence of independent repetitions of the simple experiment represented by the probabilities p_u on S. In the case where $S = \{0, 1\}$ and $p_0 = p_1 = \frac{1}{2}$, it is a model for independent tosses of a fair coin, an alternative to the model used in Section 1.

The definition (2.21) presents a consistency problem, since the cylinder A will have other representations. Suppose that A is also given by (2.19). If $n = m$, then H and H' must coincide, and there is nothing to prove. Suppose then (symmetry) that $n < m$. Then H' must consist of those $(u_1, \dots, u_m)$ in S^m for which $(u_1, \dots, u_n)$ lies in H: $H' = H \times S^{m-n}$. But then

$$(2.23) \qquad \sum_{H'} p_{u_1} \cdots p_{u_n} p_{u_{n+1}} \cdots p_{u_m} = \sum_H p_{u_1} \cdots p_{u_n} \sum_{S^{m-n}} p_{u_{n+1}} \cdots p_{u_m}$$

$$= \sum_H p_{u_1} \cdots p_{u_n}.$$

The definition (2.21) is therefore consistent. And finite additivity is now easy: Suppose that A and B are disjoint cylinders given by (2.17) and (2.18).

Suppose that $n \leq m$, and put A in the form (2.19). Since A and B are disjoint, H' and I must be disjoint as well, and by (2.20),

$$(2.24) \qquad P(A \cup B) = \sum_{H' \cup I} p_{u_1} \cdots p_{u_m} = P(A) + P(B).$$

Taking $H = S^r$ in (2.21) shows that $P(S^\infty) = 1$. Therefore, (2.21) *defines a finitely additive probability measure on the field $\mathscr{C}_0$.*

Now, P is countably additive on $\mathscr{C}_0$, but this requires no further argument, because of the following completely general result.

Theorem 2.3. *Every finitely additive probability measure on the field $\mathscr{C}_0$ of cylinders in S^∞ is in fact countably additive.*

The proof depends on this fundamental fact:

Lemma. *If $A_n \downarrow A$, where the A_n are nonempty cylinders, then A is nonempty.*

PROOF OF THEOREM 2.3. Assume that the lemma is true, and apply Example 2.10 to the measure P in question: If $A_n \downarrow \varnothing$ for sets in $\mathscr{C}_0$ (cylinders) but $P(A_n)$ does *not* converge to 0, then $P(A_n) \geq \epsilon > 0$ for some ϵ. But then the A_n are nonempty, which by the lemma makes $A_n \downarrow \varnothing$ impossible. ∎

PROOF OF THE LEMMA.[†] Suppose that A_t is a cylinder of rank m_t, say

$$(2.25) \qquad A_t = \left[\omega : \left(z_1(\omega), \ldots, z_{m_t}(\omega) \right) \in H_t \right],$$

where $H_t \subset S^{m_t}$. Choose a point ω_n in A_n, which is nonempty by assumption. Write the components of the sequences in a square array:

$$(2.26) \qquad \begin{array}{cccc} z_1(\omega_1) & z_1(\omega_2) & z_1(\omega_3) & \cdots \\ z_2(\omega_1) & z_2(\omega_2) & z_2(\omega_3) & \cdots \\ \vdots & \vdots & \vdots & \end{array}$$

The nth *column* of the array gives the components of ω_n.

Now argue by a modification of the diagonal method [A14]. Since S is finite, some element u_1 of S appears infinitely often in the first row of (2.26): for an increasing sequence $\{n_{1,k}\}$ of integers, $z_1(\omega_{n_{1,k}}) = u_1$ for all k. By the same reasoning, there exist an increasing subsequence $\{n_{2,k}\}$ of $\{n_{1,k}\}$ and an

---

[†] The lemma is a special case of Tychonov's theorem: If S is given the discrete topology, the topological product S^∞ is compact (and the cylinders are closed).

element u_2 of S such that $z_2(\omega_{n_{2,k}}) = u_2$ for all k. Continue. If $n_k = n_{k,k}$, then $z_r(\omega_{n_k}) = u_r$ for $k \geq r$, and hence $(z_1(\omega_{n_k}), \ldots, z_r(\omega_{n_k})) = (u_1, \ldots, u_r)$ for $k \geq r$.

Let ω° be the element of S^∞ with components u_r: $\omega^\circ = (u_1, u_2, \ldots) = (z_1(\omega^\circ), z_2(\omega^\circ), \ldots)$. Let t be arbitrary. If $k \geq t$, then (n_k is increasing) $n_k \geq t$ and hence $\omega_{n_k} \in A_{n_k} \subset A_t$. It follows by (2.25) that, for $k \geq t$, H_t contains the point $(z_1(\omega_{n_k}), \ldots, z_{m_t}(\omega_{n_k}))$ of S^{m_t}. But for $k \geq m_t$, this point is identical with $(z_1(\omega^\circ), \ldots, z_{m_t}(\omega^\circ))$, which therefore lies in H_t. Thus ω° is a point common to all the A_t. ∎

Let $\mathscr{C}$ be the σ-field in S^∞ generated by $\mathscr{C}_0$. By the general theory of the next section, the probability measure P defined on $\mathscr{C}_0$ by (2.21) extends to $\mathscr{C}$ The term *product measure*, properly speaking, applies to the extended P. Thus $(S^\infty, \mathscr{C}, P)$ is a probability space, one important in ergodic theory (Section 24).

Suppose that $S = \{0, 1\}$ and $p_0 = p_1 = \frac{1}{2}$. In this case, $(S^\infty, \mathscr{C}, P)$ is closely related to $((0, 1], \mathscr{B}, \lambda)$, although there are essential differences. The sequence (2.15) can end in 0's, but (1.8) cannot. Thin cylinders are like dyadic intervals, but the sets in $\mathscr{C}_0$ (the cylinders) correspond to the finite disjoint unions of intervals with dyadic endpoints, a field somewhat smaller than $\mathscr{B}_0$. While nonempty sets in $\mathscr{B}_0$ (for example, $(\frac{1}{2}, \frac{1}{2} + 2^{-n}]$) can contract to the empty set, nonempty sets in $\mathscr{C}_0$ cannot. The lemma above plays here the role the Heine-Borel theorem plays in the proof of Theorem 1.3. The product probability measure constructed here on $\mathscr{C}_0$ (in the case $S = \{0, 1\}$, $p_0 = p_1 = \frac{1}{2}$, that is) is analogous to Lebesgue measure on $\mathscr{B}_0$ But a finitely additive probability measure on $\mathscr{B}_0$ can fail to be countably additive,[†] which cannot happen in $\mathscr{C}_0$.

Constructing σ-Fields[*]

The σ-field $\sigma(\mathscr{A})$ generated by $\mathscr{A}$ was defined from above or from the outside, so to speak, by intersecting all the σ-fields that contain $\mathscr{A}$ (including the σ-field consisting of all the subsets of Ω). Can $\sigma(\mathscr{A})$ somehow be constructed from the inside by repeated finite and countable set-theoretic operations starting with sets in $\mathscr{A}$?

For any class $\mathscr{H}$ of sets in Ω let $\mathscr{H}^*$ consist of the sets in $\mathscr{H}$, the complements of sets in $\mathscr{H}$, and the finite and countable unions of sets in $\mathscr{H}$. Given a class $\mathscr{A}$, put $\mathscr{A}_0 = \mathscr{A}$ and define $\mathscr{A}_1, \mathscr{A}_2, \ldots$ inductively by

$$(2.27) \qquad\qquad \mathscr{A}_n = \mathscr{A}_{n-1}^*.$$

That each $\mathscr{A}_n$ is contained in $\sigma(\mathscr{A})$ follows by induction. One might hope that $\mathscr{A}_n = \sigma(\mathscr{A})$ for some n, or at least that $\bigcup_{n=0}^\infty \mathscr{A}_n = \sigma(\mathscr{A})$. But this process applied to the class of intervals fails to account for all the Borel sets.

Let $\mathscr{J}_0$ consist of the empty set and the intervals in $\Omega = (0, 1]$ with rational endpoints, and define $\mathscr{J}_n = \mathscr{J}_{n-1}^*$ for $n = 1, 2, \ldots$. *It will be shown that* $\bigcup_{n=0}^\infty \mathscr{J}_n$ *is strictly smaller than* $\mathscr{B} = \sigma(\mathscr{J}_0)$.

[†]See Problem 2.15.
[*]This topic may be omitted.

If a_n and b_n are rationals decreasing to a and b, then $(a, b] = \bigcup_m \bigcap_n (a_m, b_n] = \bigcup_m (\bigcup_n (a_m, b_n]^c)^c \in \mathscr{I}_4$. The result would therefore not be changed by including in $\mathscr{I}_0$ all the intervals in $(0, 1]$.

To prove $\bigcup_{n=0}^{\infty} \mathscr{I}_n$ smaller than $\mathscr{B}$, first put

$$(2.28) \qquad \Psi(A_1, A_2, \ldots) = A_1^c \cup A_2 \cup A_3 \cup A_4 \cup \cdots.$$

Since $\mathscr{I}_{n-1}$ contains $\Omega = (0, 1]$ and the empty set, every element of $\mathscr{I}_n$ has the form (2.28) for some sequence $A_1, A_2, \ldots$ of sets in $\mathscr{I}_{n-1}$. Let every positive integer appear exactly once in the square array

$$
\begin{array}{cccc}
m_{11} & m_{12} & \cdots \\
m_{21} & m_{22} & \cdots \\
\vdots & \vdots
\end{array}
$$

Inductively define

$$(2.29) \qquad \Phi_0(A_1, A_2, \ldots) = A_1,$$

$$\Phi_n(A_1, A_2, \ldots) = \Psi\Big(\Phi_{n-1}(A_{m_{11}}, A_{m_{12}}, \ldots), \Phi_{n-1}(A_{m_{21}}, A_{m_{22}}, \ldots), \ldots\Big),$$

$$n = 1, 2, \ldots.$$

It follows by induction that every element of $\mathscr{I}_n$ has the form $\Phi_n(A_1, A_2, \ldots)$ for some sequence of sets in $\mathscr{I}_0$. Finally, put

$$(2.30) \quad \Phi(A_1, A_2, \ldots) = \Phi_1(A_{m_{11}}, A_{m_{12}}, \ldots) \cup \Phi_2(A_{m_{21}}, A_{m_{22}}, \ldots) \cup \cdots.$$

Then every element of $\bigcup_{n=0}^{\infty} \mathscr{I}_n$ has the form (2.30) for some sequence $A_1, A_2, \ldots$ of sets in $\mathscr{I}_0$.

If $A_1, A_2, \ldots$ are in $\mathscr{B}$, then (2.28) is in $\mathscr{B}$; it follows by induction that each $\Phi_n(A_1, A_2, \ldots)$ is in $\mathscr{B}$ and therefore that (2.30) is in $\mathscr{B}$.

With each ω in $(0, 1]$ associate the sequence $(\omega_1, \omega_2, \ldots)$ of positive integers such that $\omega_1 + \cdots + \omega_k$ is the position of the kth 1 in the nonterminating dyadic expansion of ω (the smallest n for which $\sum_{j=1}^{n} d_j(\omega) = k$). Then $\omega \leftrightarrow (\omega_1, \omega_2, \ldots)$ is a one-to-one correspondence between $(0, 1]$ and the set of all sequences of positive integers. Let $I_1, I_2, \ldots$ be an enumeration of the sets in $\mathscr{I}_0$, put $\varphi(\omega) = \Phi(I_{\omega_1}, I_{\omega_2}, \ldots)$, and define $B = [\omega: \omega \notin \varphi(\omega)]$. It will be shown that B is a Borel set but is not contained in any of the $\mathscr{I}_n$.

Since ω lies in B if and only if ω lies outside $\varphi(\omega)$, $B \neq \varphi(\omega)$ for every ω. But every element of $\bigcup_{n=0}^{\infty} \mathscr{I}_n$ has the form (2.30) for some sequence in $\mathscr{I}_0$ and hence has the form $\varphi(\omega)$ for some ω. Therefore, B is not a member of $\bigcup_{n=0}^{\infty} \mathscr{I}_n$.

It remains to show that B is a Borel set. Let $D_k = [\omega: \omega \in I_{\omega_k}]$. Since $L_k(n) = [\omega: \omega_1 + \cdots + \omega_k = n] = [\omega: \sum_{j=1}^{n-1} d_j(\omega) < k = \sum_{j=1}^{n} d_j(\omega)]$ is a Borel set, so are $[\omega: \omega_k = n] = \bigcup_{m=1}^{\infty} L_{k-1}(m) \cap L_k(m + n)$ and

$$D_k = \Big[\omega: \omega \in I_{\omega_k}\Big] = \bigcup_n \Big([\omega: \omega_k = n] \cap I_n\Big).$$

Suppose that it is shown that

$$(2.31) \qquad \left[\omega \cdot \omega \in \Phi_n\left(I_{\omega_{u_1}}, I_{\omega_{u_2}} \ldots \right) \right] = \Phi_n\left(D_{u_1}, D_{u_2}, \ldots \right)$$

for every n and every sequence $u_1, u_2, \ldots$ of positive integers. It will then follow from the definition (2.30) that

$$B^c = \left[\omega : \omega \in \varphi(\omega) \right] = \bigcup_{n=1}^{\infty} \left[\omega : \omega \in \Phi_n\left(I_{\omega_{m_{n1}}}, I_{\omega_{m_{n2}}}, \ldots \right) \right]$$

$$= \bigcup_{n=1}^{\infty} \Phi_n\left(D_{m_{n1}}, D_{m_{n2}}, \ldots \right) = \Phi(D_1, D_2, \ldots).$$

But as remarked above, (2.30), is a Borel set if the A_n are. Therefore, (2.31) will imply that B^c and B are Borel sets.

If $n = 0$, (2.31) holds because it reduces by (2.29) to $[\omega : \omega \in I_{\omega_{u_1}}] = D_{u_1}$. Suppose that (2.31) holds with $n - 1$ in place of n. Consider the condition

$$(2.32) \qquad \omega \in \Phi_{n-1}\left(I_{\omega_{u_{m_{k_1}}}}, I_{\omega_{u_{m_{k_2}}}}, \cdot \right)$$

By (2.28) and (2.29), a necessary and sufficient condition for $\omega \in \Phi_n(I_{\omega_{u_1}}, I_{\omega_{u_2}}, \ldots)$ is that either (2.32) is false for $k = 1$ or else (2.32) is true for some k exceeding 1. But by the induction hypothesis, (2.32) and its negation can be replaced by $\omega \in \Phi_{n-1}(D_{u_{m_{k_1}}}, D_{u_{m_{k_2}}}, \ldots)$ and its negation. Therefore, $\omega \in \Phi_n(I_{\omega_{u_1}}, I_{\omega_{u_2}}, \ldots)$ if and only if $\omega \in \Phi_n(D_{u_1}, D_{u_2}, \ldots)$.

Thus $\bigcup_n \mathscr{I}_n \neq \mathscr{B}$, and there are Borel sets that cannot be arrived at from the intervals by any finite sequence of set-theoretic operations, each operation being finite or countable. It can even be shown that there are Borel sets that cannot be arrived at by any *countable* sequence of these operations. On the other hand, every Borel set can be arrived at by a countable *ordered set* of these operations if it is not required that they be performed in a simple *sequence*. The proof of this statement—and indeed even a precise explanation of its meaning—depends on the theory of infinite ordinal numbers.[†]

PROBLEMS

2.1. Define $x \vee y = \max\{x, y\}$, and for a collection $\{x_\alpha\}$ define $\bigvee_\alpha x_\alpha = \sup_\alpha x_\alpha$; define $x \wedge y = \min\{x, y\}$ and $\bigwedge_\alpha x_\alpha = \inf_\alpha x_\alpha$. Prove that $I_{A \cup B} = I_A \vee I_B$, $I_{A \cap B} = I_A \wedge I_B$, $I_{A^c} = 1 - I_A$, and $I_{A \triangle B} = |I_A - I_B|$, in the sense that there is equality at each point of Ω. Show that $A \subset B$ if and only if $I_A \leq I_B$ pointwise. Check the equation $x \wedge (y \vee z) = (x \wedge y) \vee (x \wedge z)$ and deduce the distributive law

$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$. By similar arguments prove that

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C),$$

$$A \Delta C \subset (A \Delta B) \cup (B \Delta C),$$

$$\left(\bigcup_n A_n \right)^c = \bigcap_n A_n^c,$$

$$\left(\bigcap_n A_n \right)^c = \bigcup_n A_n^c.$$

2.2. Let $A_1, \ldots, A_n$ be arbitrary events, and put $U_k = \bigcup (A_{i_1} \cap \cdots \cap A_{i_k})$ and $I_k = \bigcap (A_{i_1} \cup \cdots \cup A_{i_k})$, where the union and intersection extend over all the k-tuples satisfying $1 \le i_1 < \cdots < i_k \le n$. Show that $U_k = I_{n-k+1}$.

2.3. (a) Suppose that $\Omega \in \mathscr{F}$ and that $A, B \in \mathscr{F}$ implies $A - B = A \cap B^c \in \mathscr{F}$. Show that $\mathscr{F}$ is a field.

(b) Suppose that $\Omega \in \mathscr{F}$ and that $\mathscr{F}$ is closed under the formation of complements and finite *disjoint* unions. Show that $\mathscr{F}$ need not be a field.

2.4. Let $\mathscr{F}_1, \mathscr{F}_2, \ldots$ be classes of sets in a common space Ω.

(a) Suppose that $\mathscr{F}_n$ are fields satisfying $\mathscr{F}_n \subset \mathscr{F}_{n+1}$. Show that $\bigcup_{n=1}^{\infty} \mathscr{F}_n$ is a field.

(b) Suppose that $\mathscr{F}_n$ are σ-fields satisfying $\mathscr{F}_n \subset \mathscr{F}_{n+1}$. Show by example that $\bigcup_{n=1}^{\infty} \mathscr{F}_n$ need not be a σ-field.

2.5. The field $f(\mathscr{A})$ generated by a class $\mathscr{A}$ in Ω is defined as the intersection of all fields in Ω containing $\mathscr{A}$.

(a) Show that $f(\mathscr{A})$ is indeed a field, that $\mathscr{A} \subset f(\mathscr{A})$, and that $f(\mathscr{A})$ is minimal in the sense that if $\mathscr{G}$ is a field and $\mathscr{A} \subset \mathscr{G}$, then $f(\mathscr{A}) \subset \mathscr{G}$.

(b) Show that for nonempty $\mathscr{A}$, $f(\mathscr{A})$ is the class of sets of the form $\bigcup_{i=1}^{m} \bigcap_{j=1}^{n_i} A_{ij}$, where for each i and j either $A_{ij} \in \mathscr{A}$ or $A_{ij}^c \in \mathscr{A}$, and where the m sets $\bigcap_{j=1}^{n_i} A_{ij}$, $1 \le i \le m$, are disjoint. The sets in $f(\mathscr{A})$ can thus be explicitly presented, which is not in general true of the sets in $\sigma(\mathscr{A})$.

2.6. ↑ (a) Show that if $\mathscr{A}$ consists of the singletons, then $f(\mathscr{A})$ is the field in Example 2.3.

(b) Show that $f(\mathscr{A}) \subset \sigma(\mathscr{A})$, that $f(\mathscr{A}) = \sigma(\mathscr{A})$ if $\mathscr{A}$ is finite, and that $\sigma(f(\mathscr{A})) = \sigma(\mathscr{A})$.

(c) Show that if $\mathscr{A}$ is countable, then $f(\mathscr{A})$ is countable.

(d) Show for fields $\mathscr{F}_1$ and $\mathscr{F}_2$ that $f(\mathscr{F}_1 \cup \mathscr{F}_2)$ consists of the finite disjoint unions of sets $A_1 \cap A_2$ with $A_i \in \mathscr{F}_i$. Extend.

2.7. 2.5↑ Let H be a a set lying outside $\mathscr{F}$, where $\mathscr{F}$ is a field [or σ-field]. Show that the field [or σ-field] generated by $\mathscr{F} \cup \{H\}$ consists of sets of the form

$$(2.33) \qquad\qquad (H \cap A) \cup (H^c \cap B), \qquad A, B \in \mathscr{F}.$$

2.8. Suppose for each A in $\mathscr{A}$ that A^c is a countable union of elements of $\mathscr{A}$. The class of intervals in $(0, 1]$ has this property. Show that $\sigma(\mathscr{A})$ coincides with the smallest class over $\mathscr{A}$ that is closed under the formation of countable unions and intersections.

2.9. Show that, if $B \in \sigma(\mathscr{A})$, then there exists a countable subclass $\mathscr{A}_B$ of $\mathscr{A}$ such that $B \in \sigma(\mathscr{A}_B)$.

2.10. (a) Show that if $\sigma(\mathscr{A})$ contains every subset of Ω, then for each pair ω and ω' of distinct points in Ω there is in $\mathscr{A}$ an A such that $I_A(\omega) \neq I_A(\omega')$

(b) Show that the reverse implication holds if Ω is countable.

(c) Show by example that the reverse implication need not hold for uncountable Ω

2.11. A σ-field is *countably generated*, or *separable*, if it is generated by some countable class of sets.

(a) Show that the σ-field $\mathscr{B}$ of Borel sets is countably generated.

(b) Show that the σ-field of Example 2.4 is countably generated if and only if Ω is countable.

(c) Suppose that $\mathscr{F}_1$ and $\mathscr{F}_2$ are σ-fields, $\mathscr{F}_1 \subset \mathscr{F}_2$, and $\mathscr{F}_2$ is countably generated. Show by example that $\mathscr{F}_1$ may not be countably generated.

2.12. Show that a σ-field cannot be countably infinite—its cardinality must be finite or else at least that of the continuum. Show by example that a field can be countably infinite.

2.13. (a) Let $\mathscr{F}$ be the field consisting of the finite and the cofinite sets in an infinite Ω, and define P on $\mathscr{F}$ by taking $P(A)$ to be 0 or 1 as A is finite or cofinite. (Note that P is not well defined if Ω is finite.) Show that P is finitely additive.

(b) Show that this P is not countably additive if Ω is countably infinite.

(c) Show that this P is countably additive if Ω is uncountable.

(d) Now let $\mathscr{F}$ be the σ-field consisting of the countable and the cocountable sets in an uncountable Ω, and define P on $\mathscr{F}$ by taking $P(A)$ to be 0 or 1 as A is countable or cocountable. (Note that P is not well defined if Ω is countable.) Show that P is countably additive.

2.14. In $(0, 1]$ let $\mathscr{F}$ be the class of sets that either (i) are of the first category [A15] or (ii) have complement of the first category. Show that $\mathscr{F}$ is a σ-field. For A in $\mathscr{F}$, take $P(A)$ to be 0 in case (i) and 1 in case (ii). Show that P is countably additive.

2.15. On the field $\mathscr{B}_0$ in $(0, 1]$ define $P(A)$ to be 1 or 0 according as there does or does not exist some positive ϵ_A (depending on A) such that A contains the interval $(\frac{1}{2}, \frac{1}{2} + \epsilon_A]$. Show that P is finitely but not countably additive. No such example is possible for the field $\mathscr{C}_0$ in S^∞ (Theorem 2.3).

2.16. (a) Suppose that P is a probability measure on a field $\mathscr{F}$. Suppose that $A_t \in \mathscr{F}$ for $t > 0$, that $A_s \subset A_t$ for $s < t$, and that $A = \bigcup_{t > 0} A_t \in \mathscr{F}$. Extend Theorem 2.1(i) by showing that $P(A_t) \uparrow P(A)$ as $t \to \infty$. Show that A necessarily lies in $\mathscr{F}$ if it is a σ-field.

(b) Extend Theorem 2.1(ii) in the same way.

2.17. Suppose that P is a probability measure on a field $\mathscr{F}$, that $A_1, A_2, \ldots,$ and $A = \bigcup_n A_n$ lie in $\mathscr{F}$, and that the A_n are nearly disjoint in the sense that $P(A_m \cap A_n) = 0$ for $m \neq n$. Show that $P(A) = \sum_n P(A_n)$.

2.18. *Stochastic arithmetic.* Define a set function P_n on the class of all subsets of $\Omega = \{1, 2, \ldots\}$ by

(2.34) $$P_n(A) = \frac{1}{n}\#[m : 1 \leq m \leq n, m \in A];$$

among the first n integers, the proportion that lie in A is just $P_n(A)$. Then P_n is a discrete probability measure. The set A has *density*

(2.35) $$D(A) = \lim_n P_n(A),$$

provided this limit exists. Let $\mathscr{D}$ be the class of sets having density.

(a) Show that D is finitely but not countably additive on $\mathscr{D}$.

(b) Show that $\mathscr{D}$ contains the empty set and Ω and is closed under the formation of complements, proper differences, and finite disjoint unions, but is not closed under the formation of countable disjoint unions or of finite unions that are not disjoint.

(c) Let $\mathscr{M}$ consist of the periodic sets $M_a = [ka : k = 1, 2, \ldots]$. Observe that

(2.36) $$P_n(M_a) = \frac{1}{n}\left\lfloor \frac{n}{a} \right\rfloor \to \frac{1}{a} = D(M_a).$$

Show that the field $f(\mathscr{M})$ generated by $\mathscr{M}$ (see Problem 2.5) is contained in $\mathscr{D}$. Show that D is completely determined on $f(\mathscr{M})$ by the value it gives for each a to the event that m is divisible by a.

(d) Assume that $\sum p^{-1}$ diverges (sum over all primes; see Problem 5.20(e)) and prove that D, although finitely additive, is not countably additive on the field $f(M)$.

(e) Euler's function $\varphi(n)$ is the number of positive integers less than n and relatively prime to it. Let $p_1, \ldots, p_r$ be the distinct prime factors of n; from the inclusion-exclusion formula for the events $[m : p_i | m]$, (2.36), and the fact that the p_i divide n, deduce

(2.37) $$\frac{\varphi(n)}{n} = \prod_{p \mid n}\left(1 - \frac{1}{p}\right).$$

(f) Show for $0 \leq x \leq 1$ that $D(A) = x$ for some A.

(g) Show that D is translation invariant: If $B = [m + 1 : m \in A]$, then B has a density if and only if A does, in which case $D(A) = D(B)$.

2.19. A probability measure space $(\Omega, \mathscr{F}, P)$ is *nonatomic* if $P(A) > 0$ implies that there exists a B such that $B \subset A$ and $0 < P(B) < P(A)$ (A and B in $\mathscr{F}$, of course).

(a) Assuming the existence of Lebesgue measure λ on $\mathscr{B}$, prove that it is nonatomic.

(b) Show in the nonatomic case that $P(A) > 0$ and $\epsilon > 0$ imply that there exists a B such that $B \subset A$ and $0 < P(B) < \epsilon$.

(c) Show in the nonatomic case that $0 \le x \le P(A)$ implies that there exists a B such that $B \subset A$ and $P(B) = x$. *Hint*: Inductively define classes $\mathscr{H}_n$, numbers h_n, and sets H_n by $\mathscr{H}_0 = \{\varnothing\} = \{H_0\}$, $\mathscr{H}_n = [H: H \subset A - \bigcup_{k<n} H_k$, $P(\bigcup_{k<n} H_k) + P(H) \le x]$, $h_n = \sup[P(H) \cdot H \in \mathscr{H}_n]$, and $P(H_n) > h_n - n^{-1}$. Consider $\bigcup_k H_k$.

(d) Show in the nonatomic case that, if $p_1, p_2, \ldots$ are nonnegative and add to 1, then A can be decomposed into sets $B_1, B_2, \ldots$ such that $P(B_n) = p_n P(A)$.

2.20. Generalize the construction of product measure: For $n = 1, 2, \ldots$, let S_n be a finite space with given probabilities p_{nu}, $u \in S_n$. Let $S_1 \times S_2 \times \cdot$ be the space of sequences (2.15), where now $z_k(\omega) \in S_k$. Define P on the class of cylinders, appropriately defined, by using the product $p_{1u_1} \cdot \cdot p_{nu_n}$ on the right in (2.21) Prove P countably additive on $\mathscr{C}_0$, and extend Theorem 2.3 and its lemma to this more general setting. Show that the lemma fails if any of the S_n are infinite

2.21. (a) Suppose that $\mathscr{A} = \{A_1, A_2, \ldots\}$ is a countable partition of Ω Show (see (2.27)) that $\mathscr{A}_1 = \mathscr{A}_0^* = \mathscr{A}^*$ coincides with $\sigma(\mathscr{A})$. This is a case where $\sigma(\mathscr{A})$ can be constructed "from the inside."

(b) Show that the set of normal numbers lies in $\mathscr{I}_6$.

(c) Show that $\mathscr{H}^* = \mathscr{H}$ if and only if $\mathscr{H}$ is a σ-field. Show that $\mathscr{I}_{n-1}$ is strictly smaller than $\mathscr{I}_n$ for all n.

2.22. Extend (2.27) to infinite ordinals α by defining $\mathscr{A}_\alpha = (\bigcup_{\beta < \alpha} \mathscr{A}_\beta)^*$. Show that, if Ω is the first uncountable ordinal, then $\bigcup_{\alpha < \Omega} \mathscr{A}_\alpha = \sigma(\mathscr{A})$. Show that, if the cardinality of $\mathscr{A}$ does not exceed that of the continuum, then the same is true of $\sigma(\mathscr{A})$. Thus $\mathscr{B}$ has the power of the continuum.

2.23. ↑ Extend (2.29) to ordinals $\alpha < \Omega$ as follows. Replace the right side of (2.28) by $\bigcup_{n=1}^\infty (A_{2n-1} \cup A_{2n}^c)$. Suppose that Φ_β is defined for $\beta < \alpha$ Let $\beta_\alpha(1), \beta_\alpha(2), \ldots$ be a sequence of ordinals such that $\beta_\alpha(n) < \alpha$ and such that if $\beta < \alpha$, then $\beta = \beta_\alpha(n)$ for infinitely many even n and for infinitely many odd n; define

$$(2.38) \quad \Phi_\alpha(A_1, A_2, \ldots)$$
$$= \Psi\Big(\Phi_{\beta_\alpha(1)}(A_{m_{11}}, A_{m_{12}}, \ldots), \Phi_{\beta_\alpha(2)}(A_{m_{21}}, A_{m_{22}}, \ldots), \ldots\Big).$$

Prove by transfinite induction that (2.38) is in $\mathscr{B}$ if the A_n are, that every element of $\mathscr{I}_\alpha$ has the form (2.38) for sets A_n in $\mathscr{I}_0$, and that (2.31) holds with α in place of n. Define $\varphi_\alpha(\omega) = \Phi_\alpha(I_{\omega_1}, I_{\omega_2}, \ldots)$, and show that $B_\alpha = [\omega: \omega \notin \varphi_\alpha(\omega)]$ lies in $\mathscr{B} - \mathscr{I}_\alpha$ for $\alpha < \Omega$. Show that $\mathscr{I}_\alpha$ is strictly smaller than $\mathscr{I}_\beta$ for $\alpha < \beta \le \Omega$.

SECTION 3. EXISTENCE AND EXTENSION

The main theorem to be proved here may be compactly stated this way:

Theorem 3.1. *A probability measure on a field has a unique extension to the generated σ-field.*

In more detail the assertion is this: Suppose that P is a probability measure on a field $\mathscr{F}_0$ of subsets of Ω, and put $\mathscr{F} = \sigma(\mathscr{F}_0)$. Then there

exists a probability measure Q on $\mathscr{F}$ such that $Q(A) = P(A)$ for $A \in \mathscr{F}_0$. Further, if Q' is another probability measure on $\mathscr{F}$ such that $Q'(A) = P(A)$ for $A \in \mathscr{F}_0$, then $Q'(A) = Q(A)$ for $A \in \mathscr{F}$.

Although the measure extended to $\mathscr{F}$ is usually denoted by the same letter as the original measure on $\mathscr{F}_0$, they are really different set functions, since they have different domains of definition. The class $\mathscr{F}_0$ is only assumed finitely additive in the theorem, but the set function P on it must be assumed countably additive (since this of course follows from the conclusion of the theorem).

As shown in Theorem 2.2, λ (initially defined for intervals as length: $\lambda(I) = |I|$) extends to a probability measure on the field $\mathscr{B}_0$ of finite disjoint unions of subintervals of $(0, 1]$. By Theorem 3.1, λ extends in a unique way from $\mathscr{B}_0$ to $\mathscr{B} = \sigma(\mathscr{B}_0)$, the class of Borel sets in $(0, 1]$. The extended λ is *Lebesgue measure* on the unit interval. Theorem 3.1 has many other applications as well.

The uniqueness in Theorem 3.1 will be proved later; see Theorem 3.3. The first project is to prove that an extension does exist.

Construction of the Extension

Let P be a probability measure on a field $\mathscr{F}_0$. The construction following extends P to a class that in general is much larger than $\sigma(\mathscr{F}_0)$ but nonetheless does not in general contain all the subsets of Ω.

For each subset A of Ω, define its *outer measure* by

$$(3.1) \qquad P^*(A) = \inf \sum_n P(A_n),$$

where the infimum extends over all finite and infinite sequences $A_1, A_2, \ldots$ of $\mathscr{F}_0$-sets satisfying $A \subset \bigcup_n A_n$. If the A_n form an efficient covering of A, in the sense that they do not overlap one another very much or extend much beyond A, then $\sum_n P(A_n)$ should be a good outer approximation to the measure of A if A is indeed to have a measure assigned it at all. Thus (3.1) represents a first attempt to assign a measure to A.

Because of the rule $P(A^c) = 1 - P(A)$ for complements (see (2.6)), it is natural in approximating A from the inside to approximate the complement A^c from the outside instead and then subtract from 1:

$$(3.2) \qquad P_*(A) = 1 - P^*(A^c).$$

This, the *inner measure* of A, is a second candidate for the measure of A.[†] A plausible procedure is to assign measure to those A for which (3.1) and (3.2)

[†]An idea which seems reasonable at first is to define $P_*(A)$ as the supremum of the sums $\sum_n P(A_n)$ for disjoint sequences of $\mathscr{F}_0$-sets in A. This will not do. For example, in the case where Ω is the unit interval, $\mathscr{F}_0$ is $\mathscr{B}_0$ (Example 2.2), and P is λ as defined by (2.12), the set N of normal numbers would have inner measure 0 because it contains no nonempty elements of $\mathscr{B}_0$; in a satisfactory theory, N will have both inner and outer measure 1.

agree, and to take the common value $P^*(A) = P_*(A)$ as the measure. Since (3.1) and (3.2) agree if and only if

$$(3.3) \qquad\qquad P^*(A) + P^*(A^c) = 1,$$

the procedure would be to consider the class of A satisfying (3.3) and use $P^*(A)$ as the measure.

It turns out to be simpler to impose on A the more stringent requirement that

$$(3.4) \qquad\qquad P^*(A \cap E) + P^*(A^c \cap E) = P^*(E)$$

hold for every set E; (3.3) is the special case $E = \Omega$, because it will turn out that $P^*(\Omega) = 1$.[†] A set A is called P^*-*measurable* if (3.4) holds for all E; let $\mathcal{M}$ be the class of such sets. What will be shown is that $\mathcal{M}$ contains $\sigma(\mathcal{F}_0)$ and that the restriction of P^* to $\sigma(\mathcal{F}_0)$ is the required extension of P.

The set function P^* has four properties that will be needed:

(i) $P^*(\varnothing) = 0$;

(ii) P^* is nonnegative: $P^*(A) \geq 0$ for every $A \subset \Omega$;

(iii) P^* is monotone: $A \subset B$ implies $P^*(A) \leq P^*(B)$;

(iv) P^* is countably subadditive: $P^*(\bigcup_n A_n) \leq \sum_n P^*(A_n)$.

The others being obvious, only (iv) needs proof. For a given ϵ, choose $\mathcal{F}_0$-sets B_{nk} such that $A_n \subset \bigcup_k B_{nk}$ and $\sum_k P(B_{nk}) < P^*(A_n) + \epsilon 2^{-n}$, which is possible by the definition (3.1). Now $\bigcup_n A_n \subset \bigcup_{n,k} B_{nk}$, so that $P^*(\bigcup_n A_n) \leq \sum_{n,k} P(B_{nk}) < \sum_n P^*(A_n) + \epsilon$, and (iv) follows.[‡] Of course, (iv) implies finite subadditivity.

By definition, A lies in the class $\mathcal{M}$ of P^*-measurable sets if it splits each E in 2^Ω in such a way that P^* adds for the pieces—that is, if (3.4) holds. Because of finite subadditivity, this is equivalent to

$$(3.5) \qquad\qquad P^*(A \cap E) + P^*(A^c \cap E) \leq P^*(E).$$

Lemma 1. *The class $\mathcal{M}$ is a field.*

[†]It also turns out, after the fact, that (3.3) implies that (3.4) holds for all E anyway, see Problem 3.2.

[‡]Compare the proof on p. 9 that a countable union of negligible sets is negligible.

PROOF.   It is clear that $\Omega \in \mathcal{M}$ and that $\mathcal{M}$ is closed under complementation. Suppose that $A, B \in \mathcal{M}$ and $E \subset \Omega$. Then

$$P^*(E) = P^*(B \cap E) + P^*(B^c \cap E)$$

$$= P^*(A \cap B \cap E) + P^*(A^c \cap B \cap E)$$

$$+ P^*(A \cap B^c \cap E) + P^*(A^c \cap B^c \cap E)$$

$$\geq P^*(A \cap B \cap E)$$

$$+ P^*\big((A^c \cap B \cap E) \cup (A \cap B^c \cap E) \cup (A^c \cap B^c \cap E)\big)$$

$$= P^*\big((A \cap B) \cap E\big) + P^*\big((A \cap B)^c \cap E\big),$$

the inequality following by subadditivity. Hence[†] $A \cap B \in \mathcal{M}$, and $\mathcal{M}$ is a field.   ■

Lemma 2.   *If* $A_1, A_2, \ldots$ *is a finite or infinite sequence of disjoint* $\mathcal{M}$-*sets, then for each* $E \subset \Omega$,

$$(3.6) \qquad\qquad P^*\left(E \cap \left(\bigcup_k A_k\right)\right) = \sum_k P^*(E \cap A_k).$$

PROOF.   Consider first the case of finitely many A_k, say n of them. For $n = 1$, there is nothing to prove. In the case $n = 2$, if $A_1 \cup A_2 = \Omega$, then (3.6) is just (3.4) with A_1 (or A_2) in the role of A. If $A_1 \cup A_2$ is smaller than Ω, split $E \cap (A_1 \cup A_2)$ by A_1 and A_1^c (or by A_2 and A_2^c) and use (3.4) and disjointness.

Assume (3.6) holds for the case of $n - 1$ sets. By the case $n = 2$, together with the induction hypothesis, $P^*(E \cap (\bigcup_{k=1}^n A_k)) = P^*(E \cap (\bigcup_{k=1}^{n-1} A_k)) + P^*(E \cap A_n) = \sum_{k=1}^n P^*(E \cap A_k)$.

Thus (3.6) holds in the finite case. For the infinite case use monotonicity: $P^*(E \cap (\bigcup_{k=1}^\infty A_k)) \geq P^*(E \cap (\bigcup_{k=1}^n A_k)) = \sum_{k=1}^n P^*(E \cap A_k)$. Let $n \to \infty$, and conclude that the left side of (3.6) is greater than or equal to the right. The reverse inequality follows by countable subadditivity.   ■

Lemma 3.   *The class* $\mathcal{M}$ *is a* σ-*field, and* P^* *restricted to* $\mathcal{M}$ *is countably additive.*

PROOF.   Suppose that $A_1, A_2, \ldots$ are disjoint $\mathcal{M}$-sets with union A. Since $F_n = \bigcup_{k=1}^n A_k$ lies in the field $\mathcal{M}$, $P^*(E) = P^*(E \cap F_n) + P^*(E \cap F_n^c)$. To the

---

[†]This proof does not work if (3.4) is weakened to (3.3).

first term on the right apply (3.6), and to the second term apply monotonicity $(F_n^c \supset A^c)$: $P^*(E) \geq \sum_{k=1}^{n} P^*(E \cap A_k) + P^*(E \cap A^c)$. Let $n \to \infty$ and use (3.6) again: $P^*(E) \geq \sum_{k=1}^{\infty} P^*(E \cap A_k) + P^*(E \cap A^c) = P^*(E \cap A) + P^*(E \cap A^c)$. Hence A satisfies (3.5) and so lies in $\mathcal{M}$, which is therefore closed under the formation of countable disjoint unions.

From the fact that $\mathcal{M}$ is a field closed under the formation of countable disjoint unions it follows that $\mathcal{M}$ is a σ-field (for sets B_k in $\mathcal{M}$, let $A_1 = B_1$ and $A_k = B_k \cap B_1^c \cap \cdots \cap B_{k-1}^c$; then the A_k are disjoint $\mathcal{M}$-sets and $\bigcup_k B_k = \bigcup_k A_k \in \mathcal{M}$). The countable additivity of P^* on $\mathcal{M}$ follows from (3.6): take $E = \Omega$. ∎

Lemmas 1, 2, and 3 use only the properties (i) through (iv) of P^* derived above. The next two use the specific assumption that P^* is defined via (3.1) from a probability measure P on the field $\mathcal{F}_0$.

Lemma 4. *If P^* is defined by (3.1), then $\mathcal{F}_0 \subset \mathcal{M}$.*

PROOF. Suppose that $A \in \mathcal{F}_0$. Given E and ϵ, choose $\mathcal{F}_0$-sets A_n such that $E \subset \bigcup_n A_n$ and $\sum_n P(A_n) \leq P^*(E) + \epsilon$. The sets $B_n = A_n \cap A$ and $C_n = A_n \cap A^c$ lie in $\mathcal{F}_0$ because it is a field. Also, $E \cap A \subset \bigcup_n B_n$ and $E \cap A^c \subset \bigcup_n C_n$; by the definition of P^* and the finite additivity of P, $P^*(E \cap A) + P^*(E \cap A^c) \leq \sum_n P(B_n) + \sum_n P(C_n) = \sum_n P(A_n) \leq P^*(E) + \epsilon$. Hence $A \in \mathcal{F}_0$ implies (3.5), and so $\mathcal{F}_0 \subset \mathcal{M}$. ∎

Lemma 5. *If P^* is defined by (3.1), then*

$$(3.7) \qquad P^*(A) = P(A) \qquad \text{for } A \in \mathcal{F}_0.$$

PROOF. It is obvious from the definition (3.1) that $P^*(A) \leq P(A)$ for A in $\mathcal{F}_0$. If $A \subset \bigcup_n A_n$, where A and the A_n are in $\mathcal{F}_0$, then by the countable subadditivity and monotonicity of P on $\mathcal{F}_0$, $P(A) \leq \sum_n P(A \cap A_n) \leq \sum_n P(A_n)$. Hence (3.7). ∎

PROOF OF EXTENSION IN THEOREM 3.1. Suppose that P^* is defined via (3.1) from a (countably additive) probability measure P on the field $\mathcal{F}_0$. Let $\mathcal{F} = \sigma(\mathcal{F}_0)$. By Lemmas 3 and 4,[†]

$$\mathcal{F}_0 \subset \mathcal{F} \subset \mathcal{M} \subset 2^\Omega.$$

By (3.7), $P^*(\Omega) = P(\Omega) = 1$. By Lemma 3, P^* (which is defined on all of 2^Ω) restricted to $\mathcal{M}$ is therefore a probability measure there. And then P^* further restricted to $\mathcal{F}$ is clearly a probability measure on that class as well.

[†] In the case of Lebesgue measure, the relation is $\mathcal{B}_0 \subset \mathcal{B} \subset \mathcal{M} \subset 2^{[0,1]}$, and each of the three inclusions is strict; see Example 2.2 and Problems 3.14 and 3.21

This measure on $\mathscr{F}$ is the required extension, because by (3.7) it agrees with P on $\mathscr{F}_0$.   ∎

Uniqueness and the π-λ Theorem

To prove the extension in Theorem 3.1 is unique requires some auxiliary concepts. A class $\mathscr{P}$ of subsets of Ω is a *π-system* if it is closed under the formation of finite intersections:

(π) $A, B \in \mathscr{P}$ implies $A \cap B \in \mathscr{P}$.

A class $\mathscr{L}$ is a *λ-system* if it contains Ω and is closed under the formation of complements and of finite and countable *disjoint* unions:

(λ_1) $\Omega \in \mathscr{L}$;

(λ_2) $A \in \mathscr{L}$ implies $A^c \in \mathscr{L}$;

(λ_3) $A_1, A_2, \ldots, \in \mathscr{L}$ and $A_n \cap A_m = \varnothing$ for $m \neq n$ imply $\bigcup_n A_n \in \mathscr{L}$.

Because of the disjointness condition in (λ_3), the definition of λ-system is weaker (more inclusive) than that of σ-field. In the presence of (λ_1) and (λ_2), which imply $\varnothing \in \mathscr{L}$, the countably infinite case of (λ_3) implies the finite one.

In the presence of (λ_1) and (λ_3), (λ_2) is equivalent to the condition that $\mathscr{L}$ is closed under the formation of proper differences:

(λ_2') $A, B \in \mathscr{L}$ and $A \subset B$ imply $B - A \in \mathscr{L}$.

Suppose, in fact, that $\mathscr{L}$ satisfies (λ_2) and (λ_3). If $A, B \in \mathscr{L}$ and $A \subset B$, then $\mathscr{L}$ contains B^c, the disjoint union $A \cup B^c$, and its complement $(A \cup B^c)^c = B - A$. Hence (λ_2'). On the other hand, if $\mathscr{L}$ satisfies (λ_1) and (λ_2'), then $A \in \mathscr{L}$ implies $A^c = \Omega - A \in \mathscr{L}$. Hence (λ_2).

Although a σ-field is a λ-system, the reverse is not true (in a four-point space take $\mathscr{L}$ to consist of $\varnothing$, Ω, and the six two-point sets). But the connection is close:

Lemma 6. *A class that is both a π-system and a λ-system is a σ-field.*

PROOF. The class contains Ω by (λ_1) and is closed under the formation of complements and finite intersections by (λ_2) and (π). It is therefore a field. It is a σ-field because if it contains sets A_n, then it also contains the disjoint sets $B_n = A_n \cap A_1^c \cap \cdots \cap A_{n-1}^c$ and by (λ_3) contains $\bigcup_n A_n = \bigcup_n B_n$.   ∎

Many uniqueness arguments depend on *Dynkin's* π-λ *theorem*:

Theorem 3.2. *If $\mathcal{P}$ is a π-system and $\mathcal{L}$ is a λ-system, then $\mathcal{P} \subset \mathcal{L}$ implies $\sigma(\mathcal{P}) \subset \mathcal{L}$.*

PROOF. Let $\mathcal{L}_0$ be the λ-system generated by $\mathcal{P}$—that is, the intersection of all λ-systems containing $\mathcal{P}$. It is a λ-system, it contains $\mathcal{P}$, and it is contained in every λ-system that contains $\mathcal{P}$ (see the construction of generated σ-fields, p. 21). Thus $\mathcal{P} \subset \mathcal{L}_0 \subset \mathcal{L}$. If it can be shown that $\mathcal{L}_0$ is also a π-system, then it will follow by Lemma 6 that it is a σ-field. From the minimality of $\sigma(\mathcal{P})$ it will then follow that $\sigma(\mathcal{P}) \subset \mathcal{L}_0$, so that $\mathcal{P} \subset \sigma(\mathcal{P}) \subset \mathcal{L}_0 \subset \mathcal{L}$. Therefore, it suffices to show that $\mathcal{L}_0$ is a π-system.

For each A, let $\mathcal{L}_A$ be the class of sets B such that $A \cap B \in \mathcal{L}_0$. If A is assumed to lie in $\mathcal{P}$, or even if A is merely assumed to lie in $\mathcal{L}_0$, then $\mathcal{L}_A$ *is a λ-system*: Since $A \cap \Omega = A \in \mathcal{L}_0$ by the assumption, $\mathcal{L}_A$ satisfies (λ_1). If $B_1, B_2 \in \mathcal{L}_A$ and $B_1 \subset B_2$, then the λ-system $\mathcal{L}_0$ contains $A \cap B_1$ and $A \cap B_2$ and hence contains the proper difference $(A \cap B_2) - (A \cap B_1) = A \cap (B_2 - B_1)$, so that $\mathcal{L}_A$ contains $B_2 - B_1$: $\mathcal{L}_A$ satisfies (λ_2). If B_n are disjoint $\mathcal{L}_A$-sets, then $\mathcal{L}_0$ contains the disjoint sets $A \cap B_n$ and hence contains their union $A \cap (\bigcup_n B_n)$: $\mathcal{L}_A$ satisfies (λ_3).

If $A \in \mathcal{P}$ and $B \in \mathcal{P}$, then ($\mathcal{P}$ is a π-system) $A \cap B \in \mathcal{P} \subset \mathcal{L}_0$, or $B \in \mathcal{L}_A$. Thus $A \in \mathcal{P}$ implies $\mathcal{P} \subset \mathcal{L}_A$, and since $\mathcal{L}_A$ is a λ-system, minimality gives $\mathcal{L}_0 \subset \mathcal{L}_A$.

Thus $A \in \mathcal{P}$ implies $\mathcal{L}_0 \subset \mathcal{L}_A$, or, to put it another way, $A \in \mathcal{P}$ and $B \in \mathcal{L}_0$ together imply that $B \in \mathcal{L}_A$ and hence $A \in \mathcal{L}_B$. (The key to the proof is that $B \in \mathcal{L}_A$ if and only if $A \in \mathcal{L}_B$.) This last implication means that $B \in \mathcal{L}_0$ implies $\mathcal{P} \subset \mathcal{L}_B$. Since $\mathcal{L}_B$ is a λ-system, it follows by minimality once again that $B \in \mathcal{L}_0$ implies $\mathcal{L}_0 \subset \mathcal{L}_B$. Finally, $B \in \mathcal{L}_0$ and $C \in \mathcal{L}_0$ together imply $C \in \mathcal{L}_B$, or $B \cap C \in \mathcal{L}_0$. Therefore, $\mathcal{L}_0$ is indeed a π-system.   ∎

Since a field is certainly a π-system, the uniqueness asserted in Theorem 3.1 is a consequence of this result:

Theorem 3.3. *Suppose that P_1 and P_2 are probability measures on $\sigma(\mathcal{P})$, where $\mathcal{P}$ is a π-system. If P_1 and P_2 agree on $\mathcal{P}$, then they agree on $\sigma(\mathcal{P})$.*

PROOF. Let $\mathcal{L}$ be the class of sets A in $\sigma(\mathcal{P})$ such that $P_1(A) = P_2(A)$. Clearly $\Omega \in \mathcal{L}$. If $A \in \mathcal{L}$, then $P_1(A^c) = 1 - P_1(A) = 1 - P_2(A) = P_2(A^c)$, and hence $A^c \in \mathcal{L}$. If A_n are disjoint sets in $\mathcal{L}$, then $P_1(\bigcup_n A_n) = \sum_n P_1(A_n) = \sum_n P_2(A_n) = P_2(\bigcup_n A_n)$, and hence $\bigcup_n A_n \in \mathcal{L}$. Therefore $\mathcal{L}$ is a λ-system. Since by hypothesis $\mathcal{P} \subset \mathcal{L}$ and $\mathcal{P}$ is a π-system, the π-λ theorem gives $\sigma(\mathcal{P}) \subset \mathcal{L}$, as required.   ∎

Note that the π-λ theorem and the concept of λ-system are exactly what are needed to make this proof work: The essential property of probability measures is countable additivity, and this is a condition on countable *disjoint* unions, the only kind involved in the requirement (λ_3) in the definition of λ-system. In this, as in many applications of the π-λ theorem, $\mathscr{L} \subset \sigma(\mathscr{P})$ and therefore $\sigma(\mathscr{P}) = \mathscr{L}$, even though the relation $\sigma(\mathscr{P}) \subset \mathscr{L}$ itself suffices for the conclusion of the theorem.

Monotone Classes

A class $\mathscr{M}$ of subsets of Ω is *monotone* if it is closed under the formation of monotone unions and intersections:

(i) $A_1, A_2, \ldots \in \mathscr{M}$ and $A_n \uparrow A$ imply $A \in \mathscr{M}$;

(ii) $A_1, A_2, \ldots \in \mathscr{M}$ and $A_n \downarrow A$ imply $A \in \mathscr{M}$.

Halmos's monotone class theorem is a close relative of the π-λ theorem but will be less frequently used in this book.

Theorem 3.4. *If $\mathscr{F}_0$ is a field and $\mathscr{M}$ is a monotone class, then $\mathscr{F}_0 \subset \mathscr{M}$ implies $\sigma(\mathscr{F}_0) \subset \mathscr{M}$.*

PROOF. Let $m(\mathscr{F}_0)$ be the minimal monotone class over $\mathscr{F}_0$—the intersection of all monotone classes containing $\mathscr{F}_0$. It is enough to prove $\sigma(\mathscr{F}_0) \subset m(\mathscr{F}_0)$; this will follow if $m(\mathscr{F}_0)$ is shown to be a field, because a monotone field is a σ-field.

Consider the class $\mathscr{G} = [A: A^c \in m(\mathscr{F}_0)]$. Since $m(\mathscr{F}_0)$ is monotone, so is $\mathscr{G}$. Since $\mathscr{F}_0$ is a field, $\mathscr{F}_0 \subset \mathscr{G}$, and so $m(\mathscr{F}_0) \subset \mathscr{G}$. Hence $m(\mathscr{F}_0)$ is closed under complementation.

Define $\mathscr{G}_1$ as the class of A such that $A \cup B \in m(\mathscr{F}_0)$ for all $B \in \mathscr{F}_0$. Then $\mathscr{G}_1$ is a monotone class and $\mathscr{F}_0 \subset \mathscr{G}_1$; from the minimality of $m(\mathscr{F}_0)$ follows $m(\mathscr{F}_0) \subset \mathscr{G}_1$. Define $\mathscr{G}_2$ as the class of B such that $A \cup B \in m(\mathscr{F}_0)$ for all $A \in m(\mathscr{F}_0)$. Then $\mathscr{G}_2$ is a monotone class. Now from $m(\mathscr{F}_0) \subset \mathscr{G}_1$ it follows that $A \in m(\mathscr{F}_0)$ and $B \in \mathscr{F}_0$ together imply that $A \cup B \in m(\mathscr{F}_0)$; in other words, $B \in \mathscr{F}_0$ implies that $B \in \mathscr{G}_2$. Thus $\mathscr{F}_0 \subset \mathscr{G}_2$; by minimality, $m(\mathscr{F}_0) \subset \mathscr{G}_2$, and hence $A, B \in m(\mathscr{F}_0)$ implies that $A \cup B \in m(\mathscr{F}_0)$. ∎

Lebesgue Measure on the Unit Interval

Consider once again the unit interval $(0, 1]$ together with the field $\mathscr{B}_0$ of finite disjoint unions of subintervals (Example 2.2) and the σ-field $\mathscr{B} = \sigma(\mathscr{B}_0)$ of Borel sets in $(0, 1]$. According to Theorem 2.2, (2.12) defines a probability measure λ on $\mathscr{B}_0$. By Theorem 3.1, λ extends to $\mathscr{B}$, the extended λ being Lebesgue measure. The probability space $((0, 1], \mathscr{B}, \lambda)$ will be the basis for much of the probability theory in the remaining sections of this chapter. A few geometric properties of λ will be considered here. Since the intervals in $(0, 1]$ form a π-system generating $\mathscr{B}$, λ is the only probability measure on $\mathscr{B}$ that assigns to each interval its length as its measure.

Some Borel sets are difficult to visualize:

Example 3.1. Let $\{r_1, r_2, \ldots\}$ be an enumeration of the rationals in $(0,1)$. Suppose that ϵ is small, and choose an open interval $I_n = (a_n, b_n)$ such that $r_n \in I_n \subset (0,1)$ and $\lambda(I_n) = b_n - a_n < \epsilon 2^{-n}$. Put $A = \bigcup_{n=1}^{\infty} I_n$. By subadditivity, $0 < \lambda(A) < \epsilon$.

Since A contains all the rationals in $(0,1)$, it is dense there. Thus A is an open, dense set with measure near 0. If I is an open subinterval of $(0,1)$, then I must intersect one of the I_n, and therefore $\lambda(A \cap I) > 0$.

If $B = (0,1) - A$ then $1 - \epsilon < \lambda(B) < 1$. The set B contains no interval and is in fact nowhere dense [A15]. Despite this, B has measure nearly 1. ■

Example 3.2. There is a set defined in probability terms that has geometric properties similar to those in the preceding example. As in Section 1, let $d_n(\omega)$ be the nth digit in the dyadic expansion of ω; see (1.7). Let $A_n = [\omega \in (0,1]: d_i(\omega) = d_{n+i}(\omega) = d_{2n+i}(\omega), i = 1, \ldots, n]$, and let $A = \bigcup_{n=1}^{\infty} A_n$. Probabilistically, A corresponds to the event that in an infinite sequence of tosses of a coin, some finite initial segment is immediately duplicated twice over. From $\lambda(A_n) = 2^n \cdot 2^{-3n}$ it follows that $0 < \lambda(A) \leq \sum_{n=1}^{\infty} 2^{-2n} = \frac{1}{3}$. Again A is dense in the unit interval; its measure, less than $\frac{1}{3}$, could be made less than ϵ by requiring that some initial segment be immediately duplicated k times over with k large. ■

The outer measure (3.1) corresponding to λ on $\mathscr{B}_0$ is the infimum of the sums $\sum_n \lambda(A_n)$ for which $A_n \in \mathscr{B}_0$ and $A \subset \bigcup_n A_n$. Since each A_n is a finite disjoint union of intervals, this outer measure is

$$(3.8) \qquad \lambda^*(A) = \inf \sum_n |I_n|,$$

where the infimum extends over coverings of A by intervals I_n. The notion of negligibility in Section 1 can therefore be reformulated: A is negligible if and only if $\lambda^*(A) = 0$. For A in $\mathscr{B}$, this is the same thing as $\lambda(A) = 0$. This covers the set N of normal numbers: Since the complement N^c is negligible and lies in $\mathscr{B}$, $\lambda(N^c) = 0$. Therefore, the Borel set N itself has probability 1: $\lambda(N) = 1$.

Completeness

This is the natural place to consider completeness, although it enters into probability theory in an essential way only in connection with the study of stochastic processes in continuous time; see Sections 37 and 38.

A probability measure space $(\Omega, \mathscr{F}, P)$ is *complete* if $A \subset B$, $B \in \mathscr{F}$, and $P(B) = 0$ together imply that $A \in \mathscr{F}$ (and hence that $P(A) = 0$). If $(\Omega, \mathscr{F}, P)$ is complete, then the conditions $A \in \mathscr{F}$, $A \vartriangle A' \subset B \in \mathscr{F}$, and $P(B) = 0$ together imply that $A' \in \mathscr{F}$ and $P(A') = P(A)$.

Suppose that $(\Omega, \mathscr{F}, P)$ is an arbitrary probability space. Define P^* by (3.1) for $\mathscr{F}_0 = \mathscr{F} = \sigma(\mathscr{F}_0)$, and consider the σ-field $\mathscr{M}$ of P^*-measurable sets. The arguments leading to Theorem 3.1 show that P^* restricted to $\mathscr{M}$ is a probability measure. If $P^*(B) = 0$ and $A \subset B$, then $P^*(A \cap E) + P^*(A^c \cap E) \le P^*(B) + P^*(E) = P^*(E)$ by monotonicity, so that A satisfies (3.5) and hence lies in $\mathscr{M}$. Thus $(\Omega, \mathscr{M}, P^*)$ is a complete probability measure space. *In any probability space it is therefore possible to enlarge the σ-field and extend the measure in such a way as to get a complete space.*

Suppose that $((0, 1], \mathscr{B}, \lambda)$ is completed in this way. The sets in the completed σ-field $\mathscr{M}$ are called *Lebesgue* sets, and λ extended to $\mathscr{M}$ is still called Lebesgue measure.

Nonmeasurable Sets

There exist in $(0, 1]$ *sets that lie outside* $\mathscr{B}$. For the construction (due to Vitali) it is convenient to use addition modulo 1 in $(0, 1]$. For $x, y \in (0, 1]$ take $x \oplus y$ to be $x + y$ or $x + y - 1$ according as $x + y$ lies in $(0, 1]$ or not.[†] Put $A \oplus x = [a \oplus x \colon a \in A]$.

Let $\mathscr{L}$ be the class of Borel sets A such that $A \oplus x$ is a Borel set and $\lambda(A \oplus x) = \lambda(A)$. Then $\mathscr{L}$ is a λ-system containing the intervals, and so $\mathscr{B} \subset \mathscr{L}$ by the π-λ theorem. Thus $A \in \mathscr{B}$ implies that $A \oplus x \in \mathscr{B}$ and $\lambda(A \oplus x) = \lambda(A)$. In this sense, λ is translation-invariant.

Define x and y to be equivalent $(x \sim y)$ if $x \oplus r = y$ for some rational r in $(0, 1]$. Let H be a subset of $(0, 1]$ consisting of exactly one representative point from each equivalence class; such a set exists under the assumption of the axiom of choice [A8]. Consider now the countably many sets $H \oplus r$ for rational r

These sets are disjoint, because no two distinct points of H are equivalent. (If $H \oplus r_1$ and $H \oplus r_2$ share the point $h_1 \oplus r_1 = h_2 \oplus r_2$, then $h_1 \sim h_2$; this is impossible unless $h_1 = h_2$, in which case $r_1 = r_2$.) Each point of $(0, 1]$ lies in one of these sets, because H has a representative from each equivalence class. (If $x \sim h \in H$, then $x = h \oplus r \in H \oplus r$ for some rational r.) Thus $(0, 1] = \bigcup_r (H \oplus r)$, a countable disjoint union.

If H were in $\mathscr{B}$, it would follow that $\lambda(0, 1] = \sum_r \lambda(H \oplus r)$. This is impossible: If the value common to the $\lambda(H \oplus r)$ is 0, it leads to $1 = 0$; if the common value is positive, it leads to a convergent infinite series of identical positive terms $(a + a + \cdots < \infty$ and $a > 0)$. Thus H lies outside $\mathscr{B}$.   ∎

Two Impossibility Theorems[*]

The argument above, which uses the axiom of choice, in fact proves this: *There exists on* $2^{(0, 1]}$ *no probability measure P such that $P(A \oplus x) = P(A)$ for all $A \in 2^{(0, 1]}$ and all* $x \in (0, 1]$. In particular it is impossible to extend λ to a translation-invariant probability measure on $2^{(0, 1]}$.

[†]This amounts to working in the circle group, where the translation $y \to x \oplus y$ becomes a rotation (1 is the identity). The rationals form a subgroup, and the set H defined below contains one element from each coset.

[*]This topic may be omitted. It uses more set theory than is assumed in the rest of the book.

There is a stronger result *There exists on* $2^{(0,1]}$ *no probability measure P such that* $P\{x\} = 0$ *for each x.* Since $\lambda\{x\} = 0$, this implies that it is impossible to extend λ to $2^{(0,1]}$ at all.[†]

The proof of this second impossibility theorem requires the well-ordering principle (equivalent to the axiom of choice) and also the continuum hypothesis. Let S be the set of sequences $(s(1), s(2), \dots)$ of positive integers. Then S has the power of the continuum. (Let the nth partial sum of a sequence in S be the position of the nth 1 in the nonterminating dyadic representation of a point in $(0, 1]$; this gives a one-to-one correspondence.) By the continuum hypothesis, the elements of S can be put in a one-to-one correspondence with the set of ordinals preceding the first uncountable ordinal. Carrying the well ordering of these ordinals over to S by means of the correspondence gives to S a well-ordering relation $\leq_w$ with the property that each element has only countably many predecessors.

For s, t in S write $s \leq t$ if $s(i) \leq t(i)$ for all $i \geq 1$. Say that t *rejects* s if $t <_w s$ and $s \leq t$, this is a transitive relation. Let T be the set of unrejected elements of S. Let V_s be the set of elements that reject s, and assume it is nonempty. If t is the first element (with respect to $\leq_w$) of V_s, then $t \in T$ (if t' rejects t, then it also rejects s, and since $t' <_w t$, there is a contradiction). Therefore, if s is rejected at all, it is rejected by an element of T.

Suppose T is countable and let $t_1, t_2, \dots$ be an enumeration of its elements. If $t^*(k) = t_k(k) + 1$, then t^* is not rejected by any t_k and hence lies in T, which is impossible because it is distinct from each t_k. Thus T is uncountable and must by the continuum hypothesis have the power of $(0, 1]$.

Let x be a one-to-one map of T onto $(0, 1]$; write the image of t as x_t. Let $A_k^i = [x_t : t(i) = k]$ be the image under x of the set of t in T for which $t(i) = k$. Since $t(i)$ must have some value k, $\bigcup_{k=1}^{\infty} A_k^i = (0, 1]$. Assume that P is countably additive and choose u in S in such a way that $P(\bigcup_{k=1}^{u(i)} A_k^i) \geq 1 - 1/2^{i+1}$ for $i \geq 1$. If

$$A = \bigcap_{i=1}^{\infty} \bigcup_{k=1}^{u(i)} A_k^i = \bigcap_{i=1}^{\infty} [x_t : t(i) \leq u(i)] = [x_t : t \leq u],$$

then $P(A) > 0$. If A is shown to be countable, this will contradict the hypothesis that each singleton has probability 0.

Now, there is some t_0 in T such that $u \leq t_0$ (if $u \in T$, take $t_0 = u$, otherwise, u is rejected by some t_0 in T). If $t \leq u$ for a t in T, then $t \leq t_0$ and hence $t \leq_w t_0$ (since otherwise t_0 rejects t). This means that $[t: t \leq u]$ is contained in the countable set $[t. t \leq_w t_0]$, and A is indeed countable.

PROBLEMS

3.1. (a) In the proof of Theorem 3.1 the assumed finite additivity of P is used twice and the assumed countable additivity of P is used once. Where?

(b) Show by example that a finitely additive probability measure on a field may not be countably subadditive. Show in fact that if a finitely additive probability measure is countably subadditive, then it is necessarily countably additive as well.

[†]This refers to a countably additive extension, of course. If one is content with finite additivity, there *is* an extension to $2^{(0,1]}$; see Problem 3.8.

(c) Suppose Theorem 2.1 were weakened by strengthening its hypothesis to the assumption that $\mathscr{F}$ is a σ-field. Why would this weakened result not suffice for the proof of Theorem 3.1?

3.2. Let P be a probability measure on a field $\mathscr{F}_0$ and for every subset A of Ω define $P^*(A)$ by (3.1). Denote also by P the extension (Theorem 3.1) of P to $\mathscr{F} = \sigma(\mathscr{F}_0)$.

(a) Show that

$$(3.9) \qquad P^*(A) = \inf[P(B): A \subset B, B \in \mathscr{F}]$$

and (see (3.2))

$$(3.10) \qquad P_*(A) = \sup[P(C): C \subset A, C \in \mathscr{F}],$$

and show that the infimum and supremum are always achieved.

(b) Show that A is P^*-measurable if and only if $P_*(A) = P^*(A)$.

(c) The outer and inner measures associated with a probability measure P on a σ-field $\mathscr{F}$ are usually *defined* by (3.9) and (3.10). Show that (3.9) and (3.10) are the same as (3.1) and (3.2) with $\mathscr{F}$ in the role of $\mathscr{F}_0$.

3.3. 2.13 2.15 3.2↑ For the following examples, describe P^* as defined by (3.1) and $\mathscr{M} = \mathscr{M}(P^*)$ as defined by the requirement (3.4). Sort out the cases in which P^* fails to agree with P on $\mathscr{F}_0$ and explain why.

(a) Let $\mathscr{F}_0$ consist of the sets $\varnothing, \{1\}, \{2, 3\}$, and $\Omega = \{1, 2, 3\}$, and define probability measures P_1 and P_2 on $\mathscr{F}_0$ by $P_1\{1\} = 0$ and $P_2\{2, 3\} = 0$. Note that $\mathscr{M}(P_1^*)$ and $\mathscr{M}(P_2^*)$ differ.

(b) Suppose that Ω is countably infinite, let $\mathscr{F}_0$ be the field of finite and cofinite sets, and take $P(A)$ to be 0 or 1 as A is finite or cofinite.

(c) The same, but suppose that Ω is uncountable.

(d) Suppose that Ω is uncountable, let $\mathscr{F}_0$ consist of the countable and the cocountable sets, and take $P(A)$ to be 0 or 1 as A is countable or cocountable.

(e) The probability in Problem 2.15.

(f) Let $P(A) = I_A(\omega_0)$ for $A \in \mathscr{F}_0$, and assume $\{\omega_0\} \in \sigma(\mathscr{F}_0)$.

3.4. Let f be a strictly increasing, strictly concave function on $[0, \infty)$ satisfying $f(0) = 0$. For $A \subset (0, 1]$, define $P^*(A) = f(\lambda^*(A))$. Show that P^* is an outer measure in the sense that it satisfies $P^*(\varnothing) = 0$ and is nonnegative, monotone, and countably subadditive. Show that A lies in $\mathscr{M}$ (defined by the requirement (3.4)) if and only if $\lambda^*(A)$ or $\lambda^*(A^c)$ is 0. Show that P^* does not arise from the definition (3.1) for any probability measure P on any field $\mathscr{F}_0$.

3.5. Let Ω be the unit square $[(x, y): 0 < x, y \le 1]$, let $\mathscr{F}$ be the class of sets of the form $[(x, y): x \in A, 0 < y \le 1]$, where $A \in \mathscr{B}$, and let P have value $\lambda(A)$ at this set. Show that $(\Omega, \mathscr{F}, P)$ is a probability measure space. Show for $A = [(x, y): 0 < x \le 1, y = \frac{1}{2}]$ that $P_*(A) = 0$ and $P^*(A) = 1$.

3.6. Let P be a *finitely* additive probability measure on a field $\mathscr{F}_0$ For $A \subset \Omega$, in analogy with (3.1) define

$$(3.11) \qquad\qquad P^\circ(A) = \inf \sum_n P(A_n),$$

where now the infimum extends over all *finite* sequences of $\mathscr{F}_0$-sets A_n satisfying $A \subset \bigcup_n A_n$. (If countable coverings are allowed, everything is different. It can happen that $P^\circ(\Omega) = 0$; see Problem 3.3(e)) Let $\mathscr{M}^\circ$ be the class of sets A such that $P^\circ(E) = P^\circ(A \cap E) + P^\circ(A^c \cap E)$ for all $E \subset \Omega$.

(a) Show that $P^\circ(\varnothing) = 0$ and that P° is nonnegative, monotone, and *finitely* subadditive Using these four properties of P°, prove: Lemma 1°· $\mathscr{M}^\circ$ is a field. Lemma 2°: If $A_1, A_2, \quad$ is a *finite* sequence of disjoint $\mathscr{M}^\circ$-sets, then for each $E \subset \Omega$,

$$(3.12) \qquad\qquad P^\circ\!\left(E \cap \left(\bigcup_k A_k \right) \right) = \sum_k P^\circ(E \cap A_k).$$

Lemma 3°: P° restricted to the field $\mathscr{M}^\circ$ is *finitely* additive.

(b) Show that if P° is defined by (3.11) (finite coverings), then: Lemma 4°· $\mathscr{F}_0 \subset \mathscr{M}^\circ$. Lemma 5°: $P^\circ(A) = P(A)$ for $A \in \mathscr{F}_0$.

(c) Define $P_\circ(A) = 1 - P^\circ(A^c)$. Prove that if $E \subset A \in \mathscr{F}_0$, then

$$(3.13) \qquad\qquad P_\circ(E) = P(A) - P^\circ(A - E).$$

3.7. 2.7 3.6↑ Suppose that H lies outside the field $\mathscr{F}_0$, and let $\mathscr{F}_1$ be the field generated by $\mathscr{F}_0 \cup \{H\}$, so that $\mathscr{F}_1$ consists of the sets $(H \cap A) \cup (H^c \cap B)$ with $A, B \in \mathscr{F}_0$. The problem is to show that a finitely additive probability measure P on $\mathscr{F}_0$ has a finitely additive extension to $\mathscr{F}_1$. Define Q on $\mathscr{F}_1$ by

$$(3.14) \qquad Q((H \cap A) \cup (H^c \cap B)) = P^\circ(H \cap A) + P_\circ(H^c \cap B)$$

for $A, B \in \mathscr{F}_0$.

(a) Show that the definition is consistent.

(b) Shows that Q agrees with P on $\mathscr{F}_0$.

(c) Show that Q is finitely additive on $\mathscr{F}_1$. Show that $Q(H) = P^\circ(H)$.

(d) Define Q' by interchanging the roles of P° and $P_\circ$ on the right in (3 14). Show that Q' is another finitely additive extension of P to $\mathscr{F}_1$. The same is true of any convex combination Q'' of Q and Q'. Show that $Q''(H)$ can take any value between $P_\circ(H)$ and $P^\circ(H)$.

3.8. ↑ Use Zorn's lemma to prove a theorem of Tarski: A finitely additive probability measure on a field has a finitely additive extension to the field of all subsets of the space.

3.9. ↑ (a) Let P be a (countably additive) probability measure on a σ-field $\mathscr{F}$. Suppose that $H \notin \mathscr{F}$, and let $\mathscr{F}_1 = \sigma(\mathscr{F} \cup \{H\})$. By adapting the ideas in Problem 3.7, show that P has a countably additive extension from $\mathscr{F}$ to $\mathscr{F}_1$.

(b) It is tempting to go on and use Zorn's lemma to extend P to a completely additive probability measure on the σ-field of all subsets of Ω. Where does the obvious proof break down?

3.10. 2.17 3.2↑ As shown in the text, a probability measure space $(\Omega, \mathcal{F}, P)$ has a complete extension—that is, there exists a complete probability measure space $(\Omega, \mathcal{F}_1, P_1)$ such that $\mathcal{F} \subset \mathcal{F}_1$ and P_1 agrees with P on $\mathcal{F}$.

(a) Suppose that $(\Omega, \mathcal{F}_2, P_2)$ is a second complete extension Show by an example in a space of two points that P_1 and P_2 need not agree on the σ-field $\mathcal{F}_1 \cap \mathcal{F}_2$.

(b) There is, however, a unique minimal complete extension: Let $\mathcal{F}^+$ consist of the sets A for which there exist $\mathcal{F}$-sets B and C such that $A \vartriangle B \subset C$ and $P(C) = 0$. Show that $\mathcal{F}^+$ is a σ-field. For such a set A define $P^+(A) = P(B)$. Show that the definition is consistent, that P^+ is a probability measure on $\mathcal{F}^+$, and that $(\Omega, \mathcal{F}^+, P^+)$ is complete. Show that, if $(\Omega, \mathcal{F}_1, P_1)$ is any complete extension of $(\Omega, \mathcal{F}, P)$, then $\mathcal{F}^+ \subset \mathcal{F}_1$ and P_1 agrees with P^+ on $\mathcal{F}^+$; $(\Omega, \mathcal{F}^+, P^+)$ is the *completion* of $(\Omega, \mathcal{F}, P)$.

(c) Show that $A \in \mathcal{F}^+$ if and only if $P_*(A) = P^*(A)$, where P_* and P^* are defined by (3.9) and (3.10), and that $P^+(A) = P_*(A) = P^*(A)$ in this case. Thus the complete extension constructed in the text is exactly *the* completion.

3.11. (a) Show that a λ-system satisfies the conditions

(λ_4) $A, B \in \mathcal{L}$ and $A \cap B = \varnothing$ imply $A \cup B \in \mathcal{L}$,

(λ_5) $A_1, A_2, \ldots \in \mathcal{L}$ and $A_n \uparrow A$ imply $A \in \mathcal{L}$,

(λ_6) $A_1, A_2, \ldots \in \mathcal{L}$ and $A_n \downarrow A$ imply $A \in \mathcal{L}$.

(b) Show that $\mathcal{L}$ is a λ-system if and only if it satisfies (λ_1), (λ_2'), and (λ_5). (Sometimes these conditions, with a redundant (λ_4), are taken as the definition.)

3.12. 2.5 3.11↑ **(a)** Show that if $\mathcal{P}$ is a π-system, then the minimal λ-system over $\mathcal{P}$ coincides with $\sigma(\mathcal{P})$.

(b) Let $\mathcal{P}$ be a π-system and $\mathcal{M}$ a monotone class. Show that $\mathcal{P} \subset \mathcal{M}$ does not imply $\sigma(\mathcal{P}) \subset \mathcal{M}$.

(c) Deduce the π-λ theorem from the monotone class theorem by showing directly that, if a λ-system $\mathcal{L}$ contains a π-system $\mathcal{P}$, then $\mathcal{L}$ also contains the field generated by $\mathcal{P}$.

3.13. 2.5↑ **(a)** Suppose that $\mathcal{F}_0$ is a field and P_1 and P_2 are probability measures on $\sigma(\mathcal{F}_0)$. Show by the monotone class theorem that if P_1 and P_2 agree on $\mathcal{F}_0$, then they agree on $\sigma(\mathcal{F}_0)$.

(b) Let $\mathcal{F}_0$ be the smallest field over the π-system $\mathcal{P}$. Show by the inclusion-exclusion formula that probability measures agreeing on $\mathcal{P}$ must agree also on $\mathcal{F}_0$. Now deduce Theorem 3.3 from part (a).

3.14. 1.5 2.22↑ Prove the existence of a Lebesgue set of Lebesgue measure 0 that is not a Borel set.

3.15. 1.3 3.6 3.14↑ The *outer content* of a set A in $(0,1]$ is $c^*(A) = \inf \sum_n |I_n|$, where the infimum extends over *finite* coverings of A by intervals I_n. Thus A is

trifling in the sense of Problem 1.3 if and only if $c^*(A) = 0$. Define *inner content* by $c_*(A) = 1 - c^*(A^c)$. Show that $c_*(A) = \sup \Sigma_n |I_n|$, where the supremum extends over finite disjoint unions of intervals I_n contained in A (of course the analogue for λ_* fails) Show that $c_*(A) \leq c^*(A)$; if the two are equal, their common value is taken as the *content* $c(A)$ of A, which is then *Jordan measurable* Connect all this with Problem 3.6.

Show that $c^*(A) = c^*(A^-)$, where A^- is the closure of A (the analogue for λ^* fails).

A trifling set is Jordan measurable Find (Pioblem 3.14) a Jordan measurable set that is not a Borel set.

Show that $c_*(A) \leq \lambda_*(A) \leq \lambda^*(A) \leq c^*(A)$. What happens in this string of inequalities if A consists of the rationals in $(0, \frac{1}{2}]$ together with the irrationals in $(\frac{1}{2}, 1]$?

3.16. 1 5↑ Deduce directly by countable additivity that the Cantor set has Lebesgue measure 0.

3.17. From the fact that $\lambda(x \oplus A) = \lambda(A)$, deduce that sums and differences of normal numbers may be nonnormal.

3.18. Let H be the nonmeasurable set constructed at the end of the section.

(a) Show that, if A is a Borel set and $A \subset H$, then $\lambda(A) = 0$—that is, $\lambda_*(H) = 0$.

(b) Show that, if $\lambda^*(E) > 0$, then E contains a nonmeasurable subset.

3.19. The aim of this problem is the construction of a Borel set A in $(0, 1)$ such that $0 < \lambda(A \cap G) < \lambda(G)$ for every nonempty open set G in $(0, 1)$.

(a) It is shown in Example 3.1 how to construct a Borel set of positive Lebesgue measure that is nowhere dense. Show that every interval contains such a set.

(b) Let $\{I_n\}$ be an enumeration of the open intervals in $(0, 1)$ with rational endpoints. Construct disjoint, nowhere dense Borel sets $A_1, B_1, A_2, B_2, \ldots$ of positive Lebesgue measure such that $A_n \cup B_n \subset I_n$.

(c) Let $A = \bigcup_k A_k$. A nonempty open G in $(0, 1)$ contains some I_n. Show that $0 < \lambda(A_n) \leq \lambda(A \cap G) < \lambda(A \cap G) + \lambda(B_n) \leq \lambda(G)$.

3.20. ↑ There is no Borel set A in $(0, 1)$ such that $a\lambda(I) \leq \lambda(A \cap I) \leq b\lambda(I)$ for every open interval I in $(0, 1)$, where $0 < a \leq b < 1$. In fact prove:

(a) If $\lambda(A \cap I) \leq b\lambda(I)$ for all I and if $b < 1$, then $\lambda(A) = 0$. *Hint.* Choose an open G such that $A \subset G \subset (0, 1)$ and $\lambda(G) < b^{-1}\lambda(A)$; represent G as a disjoint union of intervals and obtain a contradiction.

(b) If $a\lambda(I) \leq \lambda(A \cap I)$ for all I and if $a > 0$, then $\lambda(A) = 1$.

3.21. Show that not every subset of the unit interval is a Lebesgue set. *Hint:* Show that λ^* is translation-invariant on $2^{(0, 1]}$; then use the first impossibility theorem (p. 45). Or use the second impossibility theorem.

SECTION 4.   DENUMERABLE PROBABILITIES

Complex probability ideas can be made clear by the systematic use of measure theory, and probabilistic ideas of extramathematical origin, such as independence, can illuminate problems of purely mathematical interest. It is to this reciprocal exchange that measure-theoretic probability owes much of its interest.

The results of this section concern infinite sequences of events in a probability space.[†] They will be illustrated by examples in the *unit interval*. By this will always be meant the triple $(\Omega, \mathscr{F}, P)$ for which Ω is $(0,1]$, $\mathscr{F}$ is the σ-field $\mathscr{B}$ of Borel sets there, and $P(A)$ is for A in $\mathscr{F}$ the Lebesgue measure $\lambda(A)$ of A. This is the space appropriate to the problems of Section 1, which will be pursued further. The definitions and theorems, as opposed to the examples, apply to *all* probability spaces. The unit interval will appear again and again in this chapter, and it is essential to keep in mind that there are many other important spaces to which the general theory will be applied later.

General Formulas

The formulas (2.5) through (2.11) will be used repeatedly. The sets involved in such formulas lie in the basic σ-field $\mathscr{F}$ by hypothesis. Any probability argument starts from given sets assumed (often tacitly) to lie in $\mathscr{F}$; further sets constructed in the course of the argument must be shown to lie in $\mathscr{F}$ as well, but it is usually quite clear how to do this.

If $P(A) > 0$, the *conditional probability* of B given A is defined in the usual way as

$$(4.1) \qquad P(B|A) = \frac{P(A \cap B)}{P(A)}.$$

There are the chain-rule formulas

$$(4.2) \qquad \begin{aligned} P(A \cap B) &= P(A)P(B|A), \\ P(A \cap B \cap C) &= P(A)P(B|A)P(C|A \cap B), \\ &\vdots \end{aligned}$$

If $A_1, A_2, \ldots$ partition Ω, then

$$(4.3) \qquad P(B) = \sum_n P(A_n)P(B|A_n).$$

[†]They come under what Borel in his first paper on the subject (see the footnote on p. 9) called *probabilités dénombrables*, hence the section heading

Note that for fixed A the function $P(B|A)$ defines a probability measure as B varies over $\mathscr{F}$.

If $P(A_n) \equiv 0$, then by subadditivity $P(\bigcup_n A_n) = 0$. If $P(A_n) \equiv 1$, then $\bigcap_n A_n$ has complement $\bigcup_n A_n^c$ of probability 0. This gives two facts used over and over again:

If $A_1, A_2, \ldots$ are sets of probability 0, so is $\bigcup_n A_n$. If $A_1, A_2, \ldots$ are sets of probability 1, so is $\bigcap_n A_n$.

Limit Sets

For a sequence $A_1, A_2, \ldots$ of sets, define a set

$$(4.4) \qquad\qquad \limsup_n A_n = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k$$

and a set

$$(4.5) \qquad\qquad \liminf_n A_n = \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k.$$

These sets[†] are the *limits superior* and *inferior* of the sequence $\{A_n\}$. They lie in $\mathscr{F}$ if all the A_n do. Now ω lies in (4.4) if and only if for each n there is some $k \geq n$ for which $\omega \in A_k$; in other words, ω lies in (4.4) if and only if it lies in *infinitely many* of the A_n. In the same way, ω lies in (4.5) if and only if there is some n such that $\omega \in A_k$ for all $k \geq n$; in other words, ω lies in (4.5) if and only if it lies in *all but finitely many* of the A_n.

Note that $\bigcap_{k=n}^{\infty} A_k \uparrow \liminf_n A_n$ and $\bigcup_{k=n}^{\infty} A_k \downarrow \limsup_n A_n$. For every m and n, $\bigcap_{k=m}^{\infty} A_k \subset \bigcup_{k=n}^{\infty} A_k$, because for $i \geq \max\{m, n\}$, A_i contains the first of these sets and is contained in the second. Taking the union over m and the intersection over n shows that (4.5) is a subset of (4.4). But this follows more easily from the observation that if ω lies in all but finitely many of the A_n then of course it lies in infinitely many of them. Facts about limits inferior and superior can usually be deduced from the logic they involve more easily than by formal set-theoretic manipulations.

If (4.4) and (4.5) are equal, write

$$(4.6) \qquad\qquad \lim_n A_n = \liminf_n A_n = \limsup_n A_n.$$

To say that A_n has limit A, written $A_n \to A$, means that the limits inferior and superior do coincide and in fact coincide with A. Since $\liminf_n A_n \subset \limsup_n A_n$ always holds, to check whether $A_n \to A$ is to check whether $\limsup_n A_n \subset A \subset \liminf_n A_n$. From $A_n \in \mathscr{F}$ and $A_n \to A$ follows $A \in \mathscr{F}$.

[†]See Problems 4 1 and 4 2 for the analogy between set-theoretic and numerical limits superior and inferior.

Example 4.1.   Consider the functions $d_n(\omega)$ defined on the unit interval by the dyadic expansion (1.7), and let $l_n(\omega)$ be the length of the run of 0's starting at $d_n(\omega)$: $l_n(\omega) = k$ if $d_n(\omega) = \cdots = d_{n+k-1}(\omega) = 0$ and $d_{n+k}(\omega) = 1$; here $l_n(\omega) = 0$ if $d_n(\omega) = 1$. Probabilities can be computed by (1.10). Since $[\omega: l_n(\omega) = k]$ is a union of 2^{n-1} disjoint intervals of length 2^{-n-k}, it lies in $\mathscr{F}$ and has probability 2^{-k-1}. Therefore, $[\omega: l_n(\omega) \geq r] = [\omega: d_i(\omega) = 0, n \leq i < n + r]$ lies also in $\mathscr{F}$ and has probability $\sum_{k \geq r} 2^{-k-1}$:

$$(4.7) \qquad\qquad P[\omega: l_n(\omega) \geq r] = 2^{-r}.$$

If A_n is the event in (4.7), then (4.4) is the set of ω such that $l_n(\omega) \geq r$ for infinitely many n, or, n being regarded as a time index, such that $l_n(\omega) \geq r$ *infinitely often*.   ■

Because of the theory of Sections 2 and 3, statements like (4.7) are valid in the sense of ordinary mathematics, and using the traditional language of probability—"neads," "runs," and so on—does not change this.

When n has the role of time, (4.4) is frequently written

$$(4.8) \qquad\qquad \limsup_n A_n = [A_n \text{ i.o.}],$$

where "i.o." stands for "infinitely often."

Theorem 4.1.   (i) *For each sequence* $\{A_n\}$,

$$(4.9) \qquad P\left(\liminf_n A_n\right) \leq \liminf_n P(A_n)$$

$$\leq \limsup_n P(A_n) \leq P\left(\limsup_n A_n\right).$$

(ii) *If* $A_n \to A$, *then* $P(A_n) \to P(A)$.

PROOF.   Clearly (ii) follows from (i). As for (i), if $B_n = \bigcap_{k=n}^{\infty} A_k$ and $C_n = \bigcup_{k=n}^{\infty} A_k$, then $B_n \uparrow \liminf_n A_n$ and $C_n \downarrow \limsup_n A_n$, so that, by parts (i) and (ii) of Theorem 2.1, $P(A_n) \geq P(B_n) \to P(\liminf_n A_n)$ and $P(A_n) \leq P(C_n) \to P(\limsup_n A_n)$.   ■

Example 4.2.   Define $l_n(\omega)$ as in Example 4.1, and let $A_n = [\omega: l_n(\omega) \geq r]$ for fixed r. By (4.7) and (4.9), $P[\omega: l_n(\omega) \geq r \text{ i.o.}] \geq 2^{-r}$. Much stronger results will be proved later.   ■

Independent Events

Events A and B are *independent* if $P(A \cap B) = P(A)P(B)$. (Sometimes an unnecessary *mutually* is put in front of *independent*.) For events of positive

probability, this is the same thing as requiring $P(B|A) = P(B)$ or $P(A|B) = P(A)$. More generally, a finite collection $A_1, \ldots, A_n$ of events is independent if

(4.10) $$P\left(A_{k_1} \cap \cdots \cap A_{k_j}\right) = P\left(A_{k_1}\right) \cdots P\left(A_{k_j}\right)$$

for $2 \le j \le n$ and $1 \le k_1 < \cdots < k_j \le n$. Reordering the sets clearly has no effect on the condition for independence, and a subcollection of independent events is also independent. An infinite (perhaps uncountable) collection of events is defined to be independent in each of its finite subcollections is.

If $n = 3$, (4.10) imposes for $j = 2$ the three constraints

(4.11) $$P(A_1 \cap A_2) = P(A_1)P(A_2), \qquad P(A_1 \cap A_3) = P(A_1)P(A_3),$$
$$P(A_2 \cap A_3) = P(A_2)P(A_3),$$

and for $j = 3$ the single constraint

(4.12) $$P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2)P(A_3).$$

Example 4.3. Consider in the unit interval the events $B_{uv} = [\omega: d_u(\omega) = d_v(\omega)]$—the uth and vth tosses agree—and let $A_1 = B_{12}$, $A_2 = B_{13}$, $A_3 = B_{23}$. Then A_1, A_2, A_3 are *pairwise* independent in the sense that (4.11) holds (the two sides of each equation being $\frac{1}{4}$). But since $A_1 \cap A_2 \subset A_3$, (4.12) does *not* hold (the left side is $\frac{1}{4}$ and the right is $\frac{1}{8}$), and the events are not independent. ∎

Example 4.4. In the discrete space $\Omega = \{1, \ldots, 6\}$ suppose each point has probability $\frac{1}{6}$ (a fair die is rolled). If $A_1 = \{1, 2, 3, 4\}$ and $A_2 = A_3 = \{4, 5, 6\}$, then (4.12) holds but none of the equations in (4.11) do. Again the events are not independent. ∎

Independence requires that (4.10) hold for each $j = 2, \ldots, n$ and each choice of $k_1, \ldots, k_j$, a total of $\sum_{j=2}^{n} \binom{n}{j} = 2^n - 1 - n$ constraints. This requirement can be stated in a different way: If $B_1, \ldots, B_n$ are sets such that for each $i = 1, \ldots, n$ either $B_i = A_i$ or $B_i = \Omega$, then

(4.13) $$P(B_1 \cap B_2 \cap \cdots \cap B_n) = P(B_1)P(B_2) \cdots P(B_n).$$

The point is that if $B_i = \Omega$, then B_i can be ignored in the intersection on the left and the factor $P(B_i) = 1$ can be ignored in the product on the right. For example, replacing A_2 by Ω reduces (4.12) to the middle equation in (4.11).

From the assumed independence of certain sets it is possible to deduce the independence of other sets.

Example 4.5. On the unit interval the events $H_n = [\omega: d_n(\omega) = 0]$, $n = 1, 2, \ldots$, are independent, the two sides of (4.10) having in this case value 2^{-j}. It seems intuitively clear that from this should follow the independence, for example, of $[\omega: d_2(\omega) = 0] = H_2$ and $[\omega: d_1(\omega) = 0, d_3(\omega) = 1] = H_1 \cap H_3^c$, since the two events involve disjoint sets of times. Further, any sets A and B depending, respectively, say, only on even and on odd times (like $[\omega: d_{2n}(\omega) = 0$ i.o.] and $[\omega: d_{2n+1}(\omega) = 1$ i.o.]) ought also to be independent. This raises the general question of what it means for A to depend only on even times. Intuitively, it requires that knowing which ones among $H_2, H_4, \ldots$ occurred entails knowing whether or not A occurred—that is, it requires that the sets $H_2, H_4, \ldots$ "determine" A. The set-theoretic form of this requirement is that A is to lie in the σ-field generated by $H_2, H_4, \ldots$. From $A \in \sigma(H_2, H_4, \ldots)$ and $B \in \sigma(H_1, H_3, \ldots)$ it ought to be possible to deduce the independence of A and B. ∎

The next theorem and its corollaries make such deductions possible. Define *classes* $\mathscr{A}_1, \ldots, \mathscr{A}_n$ in the basic σ-field $\mathscr{F}$ to be independent if for each choice of A_i from $\mathscr{A}_i$, $i = 1, \ldots, n$, the events $A_1, \ldots, A_n$ are independent. This is the same as requiring that (4.13) hold whenever for each i, $1 \leq i \leq n$, either $B_i \in \mathscr{A}_i$ or $B_i = \Omega$.

Theorem 4.2. *If $\mathscr{A}_1, \ldots, \mathscr{A}_n$ are independent and each $\mathscr{A}_i$ is a π-system, then $\sigma(\mathscr{A}_1), \ldots, \sigma(\mathscr{A}_n)$ are independent.*

PROOF. Let $\mathscr{B}_i$ be the class $\mathscr{A}_i$ augmented by Ω (which may be an element of $\mathscr{A}_i$ to start with). Then each $\mathscr{B}_i$ is a π-system, and by the hypothesis of independence, (4.13) holds if $B_i \in \mathscr{B}_i$, $i = 1, \ldots, n$. For fixed sets $B_2, \ldots, B_n$ lying respectively in $\mathscr{B}_2, \ldots, \mathscr{B}_n$, let $\mathscr{L}$ be the class of $\mathscr{F}$-sets B_1 for which (4.13) holds. Then $\mathscr{L}$ is a λ-system containing the π-system $\mathscr{B}_1$ and hence (Theorem 3.2) containing $\sigma(\mathscr{B}_1) = \sigma(\mathscr{A}_1)$. Therefore, (4.13) holds if $B_1, B_2, \ldots, B_n$ lie respectively in $\sigma(\mathscr{A}_1), \mathscr{B}_2, \ldots, \mathscr{B}_n$, which means that $\sigma(\mathscr{A}_1), \mathscr{A}_2, \ldots, \mathscr{A}_n$ are independent. Clearly the argument goes through if 1 is replaced by any of the indices $2, \ldots, n$.

From the independence of $\sigma(\mathscr{A}_1), \mathscr{A}_2, \ldots, \mathscr{A}_n$ now follows that of $\sigma(\mathscr{A}_1), \sigma(\mathscr{A}_2), \mathscr{A}_3, \ldots, \mathscr{A}_n$, and so on. ∎

If $\mathscr{A} = \{A_1, \ldots, A_k\}$ is finite, then each A in $\sigma(\mathscr{A})$ can be expressed by a "formula" such as $A = A_2 \cap A_5^c$ or $A = (A_2 \cap A_7) \cup (A_3 \cap A_7^c \cap A_8)$. If $\mathscr{A}$ is infinite, the sets in $\sigma(\mathscr{A})$ may be very complicated; the way to make precise the idea that the elements of $\mathscr{A}$ "determine" A is not to require formulas, but simply to require that A lie in $\sigma(\mathscr{A})$.

Independence for an infinite collection of classes is defined just as in the finite case: $[\mathscr{A}_\theta: \theta \in \Theta]$ is independent if the collection $[A_\theta: \theta \in \Theta]$ of sets is independent for each choice of A_θ from $\mathscr{A}_\theta$. This is equivalent to the independence of each finite subcollection $\mathscr{A}_{\theta_1}, \ldots, \mathscr{A}_{\theta_n}$ of classes, because of

the way independence for infinite classes of sets is defined in terms of independence for finite classes. Hence Theorem 4.2 has an immediate consequence:

Corollary 1. *If* $\mathscr{A}_\theta$, $\theta \in \Theta$, *are independent and each* $\mathscr{A}_\theta$ *is a* π-*system, then* $\sigma(\mathscr{A}_\theta)$, $\theta \in \Theta$, *are independent.*

Corollary 2. *Suppose that the array*

(4.14)
$$
\begin{array}{cc}
A_{11} & A_{12} & \cdots \\
A_{21} & A_{22} & \cdots \\
\vdots & \vdots
\end{array}
$$

of events is independent; here each row is a finite or infinite sequence, and there are finitely or infinitely many rows. If $\mathscr{F}_i$ *is the* σ-*field generated by the* ith *row, then* $\mathscr{F}_1, \mathscr{F}_2, \ldots$ *are independent.*

PROOF. If $\mathscr{A}_i$ is the class of all finite intersections of elements of the ith row of (4.14), then $\mathscr{A}_i$ is a π-system and $\sigma(\mathscr{A}_i) = \mathscr{F}_i$. Let I be a finite collection of indices (integers), and for each i in I let J_i be a finite collection of indices. Consider for $i \in I$ the element $C_i = \bigcap_{j \in J_i} A_{ij}$ of $\mathscr{A}_i$. Since every finite subcollection of the array (4.14) is independent (the intersections and products here extend over $i \in I$ and $j \in J_i$),

$$
P\left(\bigcap_i C_i \right) = P\left(\bigcap_i \bigcap_j A_{ij} \right) = \prod_i \prod_j P(A_{ij}) = \prod_i P(\bigcap_j A_{ij})
$$

$$
= \prod_i P(C_i).
$$

It follows that the classes $\mathscr{A}_1, \mathscr{A}_2, \ldots$ are independent, so that Corollary 1 applies. ∎

Corollary 2 implies the independence of the events discussed in Example 4.5. The array (4.14) in this case has two rows:

$$
\begin{array}{ccc}
H_2 & H_4 & H_6 & \cdots \\
H_1 & H_3 & H_5 & \cdots
\end{array}
$$

Theorem 4.2 also implies, for example, that for independent $A_1, \ldots, A_n$,

(4.15)
$$
P(A_1^c \cap \cdots \cap A_k^c \cap A_{k+1} \cap \cdots \cap A_n)
$$
$$
= P(A_1^c) \cdots P(A_k^c) P(A_{k+1}) \cdots P(A_n).
$$

To prove this, let $\mathscr{A}_i$ consist of A_i alone; of course, $A_i^c \in \sigma(\mathscr{A}_i)$. In (4.15) any subcollection of the A_i could be replaced by their complements.

Example 4.6. Consider as in Example 4.3 the events B_{uv} that, in a sequence of tosses of a fair coin, the uth and vth outcomes agree. Let $\mathscr{A}_1$ consist of the events B_{12} and B_{13}, and let $\mathscr{A}_2$ consist of the event B_{23}. Since these three events are pairwise independent, the classes $\mathscr{A}_1$ and $\mathscr{A}_2$ are independent. Since $B_{23} = (B_{12} \triangle B_{13})^c$ lies in $\sigma(\mathscr{A}_1)$, however, $\sigma(\mathscr{A}_1)$ and $\sigma(\mathscr{A}_2)$ are not independent. The trouble is that $\mathscr{A}_1$ is not a π-system, which shows that this condition in Theorem 4.2 is essential. ∎

Example 4.7. If $\mathscr{A} = \{A_1, A_2, \ldots\}$ is a finite or countable partition of Ω and $P(B|A_i) = p$ for each A_i of positive probability, then $P(B) = p$ and B is independent of $\sigma(\mathscr{A})$. If Σ' denotes summation over those i for which $P(A_i) > 0$, then $P(B) = \Sigma' P(A_i \cap B) = \Sigma' P(A_i)p = p$, and so B is independent of each set in the π-system $\mathscr{A} \cup \{\varnothing\}$. ∎

Subfields

Theorem 4.2 involves a number of σ-fields at once, which is characteristic of probability theory; measure theory not directed toward probability usually involves only one all-embracing σ-field $\mathscr{F}$. In proability, σ-fields in $\mathscr{F}$—that is, sub-σ-fields of $\mathscr{F}$—play an important role. To understand their function it helps to have an informal, intuitive way of looking at them.

A subclass $\mathscr{A}$ of $\mathscr{F}$ corresponds heuristically to *partial information*. Imagine that a point ω is drawn from Ω according to the probabilities given by P: ω lies in A with probability $P(A)$. Imagine also an observer who does not know which ω it is that has been drawn but who does know for each A in $\mathscr{A}$ whether $\omega \in A$ or $\omega \notin A$—that is, who does not know ω but does know the value of $I_A(\omega)$ for each A in $\mathscr{A}$. Identifying this partial information with the class $\mathscr{A}$ itself will illuminate the connection between various measure-theoretic concepts and the premathematical ideas lying behind them.

The set B is by definition independent of the class $\mathscr{A}$ if $P(B|A) = P(B)$ for all sets A in $\mathscr{A}$ for which $P(A) > 0$. Thus if B is independent of $\mathscr{A}$, then the observer's probability for B is $P(B)$ even after he has received the information in $\mathscr{A}$; in this case $\mathscr{A}$ contains no information about B. The point of Theorem 4.2 is that this remains true even if the observer is given the information in $\sigma(\mathscr{A})$, provided that $\mathscr{A}$ is a π-system. It is to be stressed that here *information*, like *observer* and *know*, is an informal, extramathematical term (in particular, it is not information in the technical sense of entropy).

The notion of partial information can be looked at in terms of partitions. Say that points ω and ω' are *$\mathscr{A}$-equivalent* if, for every A in $\mathscr{A}$, ω and ω' lie

either both in A or both in A^c—that is, if

$$(4.16) \qquad\qquad I_A(\omega) = I_A(\omega'), \qquad A \in \mathscr{A}.$$

This relation partitions Ω into sets of equivalent points; call this the $\mathscr{A}$ partition.

Example 4.8. If ω and ω' are $\sigma(\mathscr{A})$-equivalent, then certainly they are $\mathscr{A}$-equivalent. For fixed ω and ω', the class of A such that $I_A(\omega) = I_A(\omega')$ is a σ-field; if ω and ω' are $\mathscr{A}$-equivalent, then this σ-field contains $\mathscr{A}$ and hence $\sigma(\mathscr{A})$, so that ω and ω' are also $\sigma(\mathscr{A})$-equivalent. Thus $\mathscr{A}$-equivalence and $\sigma(\mathscr{A})$-equivalence are the same thing, and the $\mathscr{A}$-partition coincides with the $\sigma(\mathscr{A})$-partition.   ■

An observer with the information in $\sigma(\mathscr{A})$ knows, not the point ω drawn, but only the equivalence class containing it. That is exactly the information he has. In Example 4.6, it is as though an observer with the items of information in $\mathscr{A}_1$ were unable to combine them to get information about B_{23}.

Example 4.9. If $H_n = [\omega: d_n(\omega) = 0]$ as in Example 4.5, and if $\mathscr{A} = \{H_1, H_3, H_5, \ldots\}$, then ω and ω' satisfy (4.16) if and only if $d_n(\omega) = d_n(\omega')$ for all odd n. The information in $\sigma(\mathscr{A})$ is thus the set of values of $d_n(\omega)$ for n odd.   ■

One who knows that ω lies in a set A has more information about ω the smaller A is. One who knows $I_A(\omega)$ for each A in a class $\mathscr{A}$, however, has more information about ω the larger $\mathscr{A}$ is. Furthermore, to have the information in $\mathscr{A}_1$ *and* the information in $\mathscr{A}_2$ is to have the information in $\mathscr{A}_1 \cup \mathscr{A}_2$, not that in $\mathscr{A}_1 \cap \mathscr{A}_2$.

The following example points up the informal nature of this interpretation of σ-fields as information.

Example 4.10. In the unit interval $(\Omega, \mathscr{F}, P)$ let $\mathscr{G}$ be the σ-field consisting of the countable and the cocountable sets. Since $P(G)$ is 0 or 1 for each G in $\mathscr{G}$, each set H in $\mathscr{F}$ is independent of $\mathscr{G}$. But in this case the $\mathscr{G}$-partition consists of the singletons, and so the information in $\mathscr{G}$ tells the observer exactly which ω in Ω has been drawn. (i) The σ-field $\mathscr{G}$ contains *no* information about H—in the sense that H and $\mathscr{G}$ are independent. (ii) The σ-field $\mathscr{G}$ contains *all* the information about H—in the sense that it tells the observer exactly which ω was drawn.   ■

In this example, (i) and (ii) stand in apparent contradiction. But the mathematics is in (i)—H and $\mathscr{G}$ *are* independent—while (ii) only concerns heuristic interpretation. The source of the difficulty or apparent paradox here lies in the unnatural structure of the σ-field $\mathscr{G}$ rather than in any deficiency in the notion of independence.[†] The heuristic equating of σ-fields and information is helpful even though it sometimes

---

[†]See Problem 4.10 for a more extreme example

breaks down, and of course proofs are indifferent to whatever illusions and vagaries brought them into existence.

The Borel–Cantelli Lemmas

This is *the first Borel–Cantelli lemma*:

Theorem 4.3.   *If* $\sum_n P(A_n)$ *converges, then* $P(\limsup_n A_n) = 0$.

PROOF.   From $\limsup_n A_n \subset \bigcup_{k=m}^{\infty} A_k$ follows $P(\limsup_n A_n) \le P(\bigcup_{k=m}^{\infty} A_k) \le \sum_{k=m}^{\infty} P(A_k)$, and this sum tends to 0 as $m \to \infty$ if $\sum_n P(A_n)$ converges.   ∎

By Theorem 4.1, $P(A_n) \to 0$ implies that $P(\liminf_n A_n) = 0$; in Theorem 4.3 hypothesis and conclusion are both stronger.

Example 4.11.   Consider the run length $l_n(\omega)$ of Example 4.1 and a sequence $\{r_n\}$ of positive reals. *If the series* $\sum 1/2^{r_n}$ *converges, then*

$$(4.17) \qquad\qquad P[\omega \colon l_n(\omega) \ge r_n \text{ i.o.}] = 0.$$

To prove this, note that if s_n is r_n rounded up to the next integer, then by (4.7), $P[\omega \colon l_n(\omega) \ge r_n] = 2^{-s_n} \le 2^{-r_n}$. Therefore, (4.17) follows by the first Borel–Cantelli lemma.

If $r_n = (1 + \epsilon) \log_2 n$ and ϵ is positive, there is convergence because $2^{-r_n} = 1/n^{1+\epsilon}$. Thus

$$(4.18) \qquad\qquad P[\omega \colon l_n(\omega) \ge (1 + \epsilon) \log_2 n \text{ i.o.}] = 0.$$

The limit superior of the ratio $l_n(\omega)/\log_2 n$ exceeds 1 if and only if ω belongs to the set in (4.18) for some positive rational ϵ. Since the union of this countable class of sets has probability 0,

$$(4.19) \qquad\qquad P\left[\omega \colon \limsup_n \frac{l_n(\omega)}{\log_2 n} > 1 \right] = 0.$$

To put it the other way around,

$$(4.20) \qquad\qquad P\left[\omega \colon \limsup_n \frac{l_n(\omega)}{\log_2 n} \le 1 \right] = 1.$$

Technically, the probability in (4.20) refers to Lebesgue measure. Intuitively, it refers to an infinite sequence of independent tosses of a fair coin.   ∎

In this example, whether $\limsup_n l_n(\omega)/\log_2 n \le 1$ holds or not is a property of ω, and the property in fact holds for ω in an $\mathscr{F}$-set of probability

1. In such a case the property is said to hold *with probability 1*, or *almost surely*. In nonprobabilistic contexts, a property that holds for ω outside a (measurable) set of measure 0 holds *almost everywhere*, or for *almost all* ω.

Example 4.12. The preceding example has an interesting arithmetic consequence. Truncating the dyadic expansion at n gives the standard $(n-1)$-place approximation $\sum_{k=1}^{n-1}d_k(\omega)2^{-k}$ to ω; the error is between 0 and 2^{-n+1}, and the error relative to the maximum is

$$(4\ 21) \qquad e_i(\omega) = \frac{\omega - \sum_{k=1}^{n-1}d_k(\omega)2^{-k}}{2^{-n+1}} = \sum_{i=1}^{\infty}d_{n+i-1}(\omega)2^{-i},$$

which lies between 0 and 1. The binary expansion of $e_n(\omega)$ begins with $l_n(\omega)$ 0's, and then comes a 1 Hence $.0\ .\ 01 \le e_n(\omega) \le .0\ldots0111\ldots$, where there are $l_n(\omega)$ 0's in the extreme terms. Therefore,

$$(4.22) \qquad \frac{1}{2^{l_n(\omega)+1}} \le e_n(\omega) \le \frac{1}{2^{l_n(\omega)}},$$

so that results on run length give information about the error of approximation.

By the left-hand inequality in (4.22), $e_n(\omega) \le x_n$ (assume that $0 < x_n \le 1$) implies that $l_n(\omega) \ge -\log_2 x_n - 1$; since $\sum 2^{-r_n} < \infty$ implies (4.17), $\sum x_n < \infty$ implies $P[\omega: e_n(\omega) \le x_n \text{ i.o.}] = 0$. (Clearly, $[\omega: e_n(\omega) \le x]$ is a Borel set) In particular,

$$(4.23) \qquad P\big[\omega: e_n(\omega) \le 1/n^{1+\epsilon} \text{ i.o.}\big] = 0.$$

Technically, this probability refers to Lebesgue measure; intuitively, it refers to a point drawn at random from the unit interval. ∎

Example 4.13. The final step in the proof of the normal number theorem (Theorem 1.2) was a disguised application of the first Borel–Cantelli lemma. If $A_n = [\omega: |n^{-1}s_n(\omega)| \ge n^{-1/8}]$, then $\sum P(A_n) < \infty$, as follows by (1.29), and so $P[A_n \text{ i.o.}] = 0$. But for ω in the set complementary to $[A_n \text{ i.o.}]$, $n^{-1}s_n(\omega) \to 0$.

The proof of Theorem 1.6 is also, in effect, an application of the first Borel–Cantelli lemma. ∎

This is *the second Borel–Cantelli lemma*:

Theorem 4.4. *If $\{A_n\}$ is an independent sequence of events and $\sum_n P(A_n)$ diverges, then $P(\limsup_n A_n) = 1$.*

PROOF. It is enough to prove that $P(\bigcup_{n=1}^{\infty}\bigcap_{k=n}^{\infty}A_k^c) = 0$ and hence enough to prove that $P(\bigcap_{k=n}^{\infty}A_k^c) = 0$ for all n. Since $1 - x \le e^{-x}$,

$$P\left(\bigcap_{k=n}^{n+j}A_k^c\right) = \prod_{k=n}^{n+j}(1 - P(A_k)) \le \exp\left[-\sum_{k=n}^{n+j}P(A_k)\right].$$

Since $\Sigma_k P(A_k)$ diverges, the last expression tends to 0 as $j \to \infty$, and hence $P(\bigcap_{k=n}^{\infty} A_k^c) = \lim_j P(\bigcap_{k=n}^{n+j} A_k^c) = 0$.   ∎

By Theorem 4.1, $\limsup_n P(A_n) > 0$ implies $P(\limsup_n A_n) > 0$; in Theorem 4.4, the hypothesis $\Sigma_n P(A_n) = \infty$ is weaker but the conclusion is stronger because of the additional hypothesis of independence.

Example 4.14.   Since the events $[\omega: l_n(\omega) = 0] = [\omega: d_n(\omega) = 1]$, $n = 1, 2, \ldots$, are independent and have probability $\frac{1}{2}$, $P[\omega: l_n(\omega) = 0 \text{ i.o.}] = 1$.

Since the events $A_n = [\omega: l_n(\omega) = 1] = [\omega: d_n(\omega) = 0, d_{n+1}(\omega) = 1]$, $n = 1, 2, \ldots$, are not independent, this argument is insufficient to prove that

(4.24) $P[\omega: l_n(\omega) = 1 \text{ i.o.}] = 1$.

But the events $A_2, A_4, A_6, \ldots$ are independent (Theorem 4.2) and their probabilities form a divergent series, and so $P[\omega: l_{2n}(\omega) = 1 \text{ i.o.}] = 1$, which implies (4.24).   ∎

Significant applications of the second Borel–Cantelli lemma usually require, in order to get around problems of dependence, some device of the kind used in the preceding example.

Example 4.15.   There is a complement to (4.17): *If r_n is nondecreasing and $\Sigma 2^{-r_n}/r_n$ diverges, then*

(4.25) $P[\omega: l_n(\omega) \geq r_n \text{ i.o.}] = 1$.

To prove this, note first that if r_n is rounded up to the next integer, then $\Sigma 2^{-r_n}/r_n$ still diverges and (4.25) is unchanged. Assume then that $r_n = r(n)$ is integral, and define $\{n_k\}$ inductively by $n_1 = 1$ and $n_{k+1} = n_k + r_{n_k}$, $k \geq 1$. Let $A_k = [\omega: l_{n_k}(\omega) \geq r_{n_k}] = [\omega: d_i(\omega) = 0, \ n_k \leq i < n_{k+1}]$; since the A_k involve nonoverlapping sequences of time indices, it follows by Corollary 2 to Theorem 4.2 that $A_1, A_2, \ldots$ are independent. By the second Borel–Cantelli lemma, $P[A_k \text{ i.o.}] = 1$ if $\Sigma_k P(A_k) = \Sigma_k 2^{-r(n_k)}$ diverges. But since r_n is nondecreasing,

$$\sum_{k \geq 1} 2^{-r(n_k)} = \sum_{k \geq 1} 2^{-r(n_k)} r^{-1}(n_k)(n_{k+1} - n_k)$$

$$\geq \sum_{k \geq 1} \sum_{n_k \leq n < n_{k+1}} 2^{-r_n} r_n^{-1} = \sum_{n \geq 1} 2^{-r_n} r_n^{-1}.$$

Thus the divergence of $\Sigma_n 2^{-r_n} r_n^{-1}$ implies that of $\Sigma_k 2^{-r(n_k)}$, and it follows that, with probability 1, $l_{n_k}(\omega) \geq r_{n_k}$ for infinitely many values of k. But this is stronger than (4.25).

The result in Example 4.2 follows if $r_n \equiv r$, but this is trivial. If $r_n = \log_2 n$, then $\Sigma 2^{-r_n}/r_n = \Sigma 1/(n \log_2 n)$ diverges, and therefore

(4.26) $P[\omega: l_n(\omega) \geq \log_2 n \text{ i.o.}] = 1$.

By (4.26) and (4.20),

$$(4.27) \qquad P\left[\omega: \limsup_n \frac{l_n(\omega)}{\log_2 n} = 1\right] = 1.$$

Thus for ω in a set of probability 1, $\log_2 n$ as a function of n is a kind of "upper envelope" for the function $l_n(\omega)$. ∎

Example 4.16. By the right-hand inequality in (4.22), if $l_n(\omega) \geq \log_2 n$, then $e_n(\omega) \leq 1/n$. Hence (4.26) gives

$$(4.28) \qquad P\left[\omega: e_n(\omega) \leq \frac{1}{n} \text{ i.o.}\right] = 1.$$

This and (4.23) show that, with probability 1, $e_n(\omega)$ has $1/n$ as a "lower envelope." The discrepancy between ω and its $(n-1)$-place approximation $\sum_{k=1}^{n-1} d_k(\omega) 2^{-k}$ will fall infinitely often below $1/(n 2^{n-1})$ but not infinitely often below $1/(n^{1+\epsilon} 2^{n-1})$. ∎

Example 4.17. Examples 4.12 and 4.16 have to do with the approximation of real numbers by rationals: Diophantine approximation. Change the $x_n = 1/n^{1+\epsilon}$ of (4.23) to $1/((n-1)\log 2)^{1+\epsilon}$. Then $\sum x_n$ still converges, and hence

$$P\left[\omega: e_n(\omega) \leq 1/\left(\log 2^{n-1}\right)^{1+\epsilon} \text{ i.o.}\right] = 0.$$

And by (4.28),

$$P\left[\omega: e_n(\omega) < 1/\log 2^{n-1} \text{ i.o.}\right] = 1.$$

The dyadic rational $\sum_{k=1}^{n-1} d_k(\omega) 2^{-k} = p/q$ has denominator $q = 2^{n-1}$, and $e_n(\omega) = q(\omega - p/q)$. There is therefore probability 1 that, if q is restricted to the powers of 2, then $0 \leq \omega - p/q < 1/(q \log q)$ holds for infinitely many p/q but $0 \leq \omega - p/q \leq 1/(q \log^{1+\epsilon} q)$ holds only for finitely many.[†] But contrast this with Theorems 1.5 and 1.6: The sharp rational approximations to a real number come not from truncating its dyadic (or decimal) expansion, but from truncating its continued-fraction expansion; see Section 24. ∎

The Zero–One Law

For a sequence $A_1, A_2, \ldots$ of events in a probability space $(\Omega, \mathscr{F}, P)$ consider the σ-fields $\sigma(A_n, A_{n+1}, \ldots)$ and their intersection

$$(4.29) \qquad \mathscr{T} = \bigcap_{n=1}^{\infty} \sigma(A_n, A_{n+1}, \ldots).$$

[†]This ignores the possibility of even p (reducible p/q); but see Problem 1.11(b). And rounding ω up to $(p+1)/q$ instead of down to p/q changes nothing; see Problem 4.13.

This is the *tail σ-field* associated with the sequence $\{A_n\}$, and its elements are called *tail events*. The idea is that a tail event is determined solely by the A_n for arbitrarily large n.

Example 4.18. Since $\limsup_m A_m = \bigcap_{k \geq n} \bigcup_{i \geq k} A_i$ and $\liminf_m A_m = \bigcup_{k \geq n} \bigcap_{i \geq k} A_i$ are both in $\sigma(A_n, A_{n+1}, \ldots)$, the limits superior and inferior are tail events for the sequence $\{A_n\}$.   ∎

Example 4.19. Let $l_n(\omega)$ be the run length, as before, and let $H_n = [\omega: d_n(\omega) = 0]$. For each n_0,

$$[\omega: l_n(\omega) \geq r_n \text{ i.o.}] = \bigcap_{n \geq n_0} \bigcup_{k \geq n} [\omega: l_k(\omega) \geq r_k]$$

$$= \bigcap_{n \geq n_0} \bigcup_{k \geq n} H_k \cap H_{k+1} \cap \cdots \cap H_{k+r_k-1}.$$

Thus $[\omega: l_n(\omega) \geq r_n \text{ i.o.}]$ is a tail event for the sequence $\{H_n\}$.   ∎

The probabilities of tail events are governed by *Kolmogorov's zero–one law*:[†]

Theorem 4.5. *If $A_1, A_2, \ldots$ is an independent sequence of events, then for each event A in the tail σ-field* (4.29), *$P(A)$ is either 0 or 1.*

PROOF. By Corollary 2 to Theorem 4.2, $\sigma(A_1), \ldots, \sigma(A_{n-1})$, $\sigma(A_n, A_{n+1}, \ldots)$ are independent. If $A \in \mathcal{T}$, then $A \in \sigma(A_n, A_{n+1}, \ldots)$ and therefore $A_1, \ldots, A_{n-1}, A$ are independent. Since independence of a collection of events is defined by independence of each finite subcollection, the sequence $A, A_1, A_2, \ldots$ is independent. By a second application of Corollary 2 to Theorem 4.2, $\sigma(A)$ and $\sigma(A_1, A_2, \ldots)$ are independent. But $A \in \mathcal{T} \subset \sigma(A_1, A_2, \ldots)$; from $A \in \sigma(A)$ and $A \in \sigma(A_1, A_2, \ldots)$ it follows that A is independent of itself: $P(A \cap A) = P(A)P(A)$. This is the same as $P(A) = (P(A))^2$ and can hold only if $P(A)$ is 0 or 1.   ∎

Example 4.20. By the zero–one law and Example 4.18, $P(\limsup_n A_n)$ is 0 or 1 if the A_n are independent. The Borel–Cantelli lemmas in this case go further and give a specific criterion in terms of the convergence or divergence of $\sum P(A_n)$.   ∎

Kolmogorov's result is surprisingly general, and it is in many cases quite easy to use it to show that the probability of some set must have one of the extreme values 0 and 1. It is perhaps curious that it should so often be very difficult to determine which of these extreme values is the right one.

[†]For a more general version, see Theorem 22.3

Example 4.21. By Kolmogorov's theorem and Example 4.19, $[\omega: l_n(\omega) \geq r_n$ i.o.] has probability 0 or 1. Call the sequence $\{r_n\}$ an *outer boundary* or an *inner boundary* according as this probability is 0 or 1.

In Example 4.11 it was shown that $\{r_n\}$ is an outer boundary if $\Sigma 2^{-r_n} < \infty$. In Example 4.15 it was shown that $\{r_n\}$ is an inner boundary if r_n is nondecreasing and $\Sigma 2^{-r_n} r_n^{-1} = \infty$. By these criteria $r_n = \theta \log_2 n$ gives an outer boundary if $\theta > 1$ and an inner boundary if $\theta \leq 1$.

What about the sequence $r_n = \log_2 n + \theta \log_2 \log_2 n$? Here $\Sigma 2^{-r_n} = \Sigma 1/n (\log_2 n)^\theta$, and this converges for $\theta > 1$, which gives an outer boundary. Now $2^{-r_n} r_n^{-1}$ is of the order $1/n (\log_2 n)^{1+\theta}$, and this diverges if $\theta \leq 0$, which gives an inner boundary (this follows indeed from (4.26)). But this analysis leaves the range $0 < \theta \leq 1$ unresolved, although every sequence is either an inner or an outer boundary This question is pursued further in Example 6.5.

■

PROBLEMS

4.1. 2.1↑ The limits superior and inferior of a numerical sequence $\{x_n\}$ can be defined as the supremum and infimum of the set of limit points—that is, the set of limits of convergent subsequences. This is the same thing as defining

$$(4.30) \qquad \limsup_n x_n = \bigwedge_{n=1}^{\infty} \bigvee_{k=n}^{\infty} x_k$$

and

$$(4.31) \qquad \liminf_n x_n = \bigvee_{n=1}^{\infty} \bigwedge_{k=n}^{\infty} x_k.$$

Compare these relations with (4.4) and (4.5) and prove that

$$I_{\limsup_n A_n} = \limsup_n I_{A_n}, \qquad I_{\liminf_n A_n} = \liminf_n I_{A_n}.$$

Prove that $\lim_n A_n$ exists in the sense of (4.6) if and only if $\lim_n I_{A_n}(\omega)$ exists for each ω.

4.2. ↑ (a) Prove that

$$\left(\limsup_n A_n\right) \cap \left(\limsup_n B_n\right) \supset \limsup_n (A_n \cap B_n),$$

$$\left(\limsup_n A_n\right) \cup \left(\limsup_n B_n\right) = \limsup_n (A_n \cup B_n),$$

$$\left(\liminf_n A_n\right) \cap \left(\liminf_n B_n\right) = \liminf_n (A_n \cap B_n),$$

$$\left(\liminf_n A_n\right) \cup \left(\liminf_n B_n\right) \subset \liminf_n (A_n \cup B_n).$$

Show by example that the two inclusions can be strict.

(b) The numerical analogue of the first of the relations in part (a) is

$$\left(\lim_{n} \sup x_n\right) \wedge \left(\lim_{n} \sup y_n\right) \geq \lim_{n} \sup\left(x_n \wedge y_n\right).$$

Write out and verify the numerical analogues of the others.

(c) Show that

$$\lim_{n} \sup A_n^c = \left(\lim_{n} \inf A_n\right)^c,$$

$$\lim_{n} \inf A_n^c = \left(\lim_{n} \sup A_n\right)^c,$$

$$\lim_{n} \sup A_n - \lim_{n} \inf A_n = \lim_{n} \sup\left(A_n \cap A_{n+1}^c\right)$$

$$= \lim_{n} \sup\left(A_n^c \cap A_{n+1}\right)$$

(d) Show that $A_n \to A$ and $B_n \to B$ together imply that $A_n \cup B_n \to A \cup B$ and $A_n \cap B_n \to A \cap B$.

4.3. Let A_n be the square $[(x, y): |x| \leq 1, |y| \leq 1]$ rotated through the angle $2\pi n\theta$. Give geometric descriptions of $\lim\sup_n A_n$ and $\lim\inf A_n$ in case

(a) $\theta = \frac{1}{8}$;

(b) θ is rational;

(c) θ is irrational. *Hint:* The $2\pi n\theta$ reduced modulo 2π are dense in $[0, 2\pi]$ if θ is irrational.

(d) When is there convergence is the sense of (4.6)?

4.4. Find a sequence for which all three inequalities in (4.9) are strict.

4.5. (a) Show that $\lim_n P(\lim\inf_k A_n \cap A_k^c) = 0$. *Hint:* Show that $\lim\sup_n \lim\inf_k A_n \cap A_k^c$ is empty.
   Put $A^* = \lim\sup_n A_n$ and $A_* = \lim\inf_n A_n$.

(b) Show that $P(A_n - A^*) \to 0$ and $P(A_* - A_n) \to 0$.

(c) Show that $A_n \to A$ (in the sense that $A = A^* = A_*$) implies $P(A \triangle A_n) \to 0$.

(d) Suppose that A_n converges to A in the weaker sense that $P(A \triangle A^*) = P(A \triangle A_*) = 0$ (which implies that $P(A^* - A_*) = 0$). Show that $P(A \triangle A_n) \to 0$ (which implies that $P(A_n) \to P(A)$).

4.6. In a space of six equally likely points (a die is rolled) find three events that are not independent even though each is independent of the intersection of the other two.

4.7. For events $A_1, \ldots, A_n$, consider the 2^n equations $P(B_1 \cap \cdots \cap B_n) = P(B_1) \cdots P(B_n)$ with $B_i = A_i$ or $B_i = A_i^c$ for each i. Show that $A_1, \ldots, A_n$ are independent if all these equations hold.

4.8. For each of the following classes $\mathscr{A}$, describe the $\mathscr{A}$-partition defined by (4.16).

(a) The class of finite and cofinite sets.

(b) The class of countable and cocountable sets.

(c) A partition (of arbitrary cardinality) of Ω.

(d) The level sets of $\sin x$ ($\Omega = R^1$).

(e) The σ-field in Problem 3.5.

4.9. 2.9 2.10↑ In connection with Example 4.8 and Problem 2 10, prove these facts:

(a) Every set in $\sigma(\mathscr{A})$ is a union of $\mathscr{A}$-equivalence classes.

(b) If $\mathscr{A} = [A_\theta : \theta \in \Theta]$, then the $\mathscr{A}$-equivalence classes have the form $\bigcap_\theta B_\theta$, where for each θ, B_θ is A_θ or A_θ^c.

(c) Every finite σ-field is generated by a finite partition of Ω.

(d) If $\mathscr{F}_0$ is a field, then each singleton, even each finite set, in $\sigma(\mathscr{F}_0)$ is a countable intersection of $\mathscr{F}_0$-sets.

4.10. 3.2↑ There is in the unit interval a set H that is nonmeasurable in the extreme sense that its inner and outer Lebesgue measures are 0 and 1 (see (3.9) and (3.10)): $\lambda_*(H) = 0$ and $\lambda^*(H) = 1$. See Problem 12.4 for the construction

Let $\Omega = (0, 1]$, let $\mathscr{G}$ consist of the Borel sets in Ω, and let H be the set just described. Show that the class $\mathscr{F}$ of sets of the form $(H \cap G_1) \cup (H^c \cap G_2)$ for G_1 and G_2 in $\mathscr{G}$ is a σ-field and that $P[(H \cap G_1) \cup (H^c \cap G_2)] = \frac{1}{2}\lambda(G_1) + \frac{1}{2}\lambda(G_2)$ consistently defines a probability measure on $\mathscr{F}$. Show that $P(H) = \frac{1}{2}$ and that $P(G) = \lambda(G)$ for $G \in \mathscr{G}$. Show that $\mathscr{G}$ is generated by a countable subclass (see Problem 2.11). Show that $\mathscr{G}$ contains all the singletons and that H and $\mathscr{G}$ are independent.

The construction proves this: *There exist a probability space $(\Omega, \mathscr{F}, P)$, a σ-field $\mathscr{G}$ in $\mathscr{F}$, and a set H in $\mathscr{F}$, such that $P(H) = \frac{1}{2}$, H and $\mathscr{G}$ are independent, and $\mathscr{G}$ is generated by a countable subclass and contains all the singletons.*

Example 4.10 is somewhat similar, but there the σ-field $\mathscr{G}$ is not countably generated and each set in it has probability either 0 or 1. In the present example $\mathscr{G}$ is countably generated and $P(G)$ assumes every value between 0 and 1 as G ranges over $\mathscr{G}$. Example 4.10 is to some extent unnatural because the $\mathscr{G}$ there is not countably generated. The present example, on the other hand, involves the pathological set H. This example is used in Section 33 in connection with conditional probability; see Problem 33.11.

4.11. (a) If $A_1, A_2, \ldots$ are independent events, then $P(\bigcap_{n=1}^\infty A_n) = \prod_{n=1}^\infty P(A_n)$ and $P(\bigcup_{n=1}^\infty A_n) = 1 - \prod_{n=1}^\infty (1 - P(A_n))$. Prove these facts and from them derive the second Borel–Cantelli lemma by the well-known relation between infinite series and products.

(b) Show that $P(\limsup_n A_n) = 1$ if for each k the series $\sum_{n > k} P(A_n | A_k^c \cap \cdots \cap A_{n-1}^c)$ diverges. From this deduce the second Borel–Cantelli lemma once again.

(c) Show by example that $P(\limsup_n A_n) = 1$ does not follow from the divergence of $\sum_n P(A_n | A_1^c \cap \cdots \cap A_{n-1}^c)$ alone.

(d) Show that $P(\limsup_n A_n) = 1$ if and only if $\sum_n P(A \cap A_n)$ diverges for each A of positive probability.

(e) If sets A_n are independent and $P(A_n) < 1$ for all n, then $P[A_n \text{ i.o.}] = 1$ if and only if $P(\bigcup_n A_n) = 1$.

4.12. (a) Show (see Example 4.21) that $\log_2 n + \log_2 \log_2 n + \theta \log_2 \log_2 \log_2 n$ is an outer boundary if $\theta > 1$. Generalize.

(b) Show that $\log_2 n + \log_2 \log_2 \log_2 n$ is an inner boundary.

4.13. Let φ be a positive function of integers, and define B_φ as the set of x in $(0,1)$ such that $|x - p/2^i| < 1/2^i\varphi(2^i)$ holds for infinitely many pairs p, i. Adapting the proof of Theorem 1.6, show directly (without reference to Example 4.12) that $\Sigma_i 1/\varphi(2^i) < \infty$ implies $\lambda(B_\varphi) = 0$.

4.14. 2.19↑   Suppose that there are in $(\Omega, \mathcal{F}, P)$ independent events $A_1, A_2, \ldots$ such that, if $\alpha_n = \min\{P(A_n), 1 - P(A_n)\}$, then $\Sigma\alpha_n = \infty$. Show that P is nonatomic.

4.15. 2.18↑   Let F be the set of square-free integers—those integers not divisible by any perfect square. Let F_l be the set of m such that $p^2 \mid m$ for no $p \le l$, and show that $D(F_l) = \prod_{p \le l}(1 - p^{-2})$. Show that $P_n(F_l - F) \le \Sigma_{p > l} p^{-2}$, and conclude that the square-free integers have density $\prod_p(1 - p^{-2}) = 6/\pi^2$.

4.16. 2.18↑   Reconsider Problem 2.18(d). If D were countably additive on $f(\mathcal{A})$, it would extend to $\sigma(\mathcal{A})$. Use the second Borel–Cantelli lemma.

SECTION 5. SIMPLE RANDOM VARIABLES

Definition

Let $(\Omega, \mathcal{F}, P)$ be an arbitrary probability space, and let X be a real-valued function on Ω; X is a *simple random variable* if it has finite range (assumes only finitely many values) and if

$$(5.1) \qquad\qquad [\omega: X(\omega) = x] \in \mathcal{F}$$

for each real x. (Of course, $[\omega: X(\omega) = x] = \varnothing \in \mathcal{F}$ for x outside the range of X.) Whether or not X satisfies this condition depends only on $\mathcal{F}$, not on P, but the point of the definition is to ensure that the probabilities $P[\omega: X(\omega) = x]$ are defined. Later sections will treat the theory of general random variables, of functions on Ω having arbitrary range; (5.1) will require modification in the general case.

The $d_n(\omega)$ of the preceding section (the digits of the dyadic expansion) are simple random variables on the unit interval: the sets $[\omega: d_n(\omega) = 0]$ and $[\omega: d_n(\omega) = 1]$ are finite unions of subintervals and hence lie in the σ-field $\mathcal{B}$ of Borel sets in $(0, 1]$. The Rademacher functions are also simple random variables. Although the concept itself is thus not entirely new, to proceed further in probability requires a systematic theory of random variables and their expected values.

The run lengths $l_n(\omega)$ satisfy (5.1) but are not simple random variables, because they have infinite range (they come under the general theory). In a discrete space, $\mathcal{F}$ consists of all subsets of Ω, so that (5.1) always holds.

It is customary in probability theory to omit the argument ω. Thus X stands for a general value $X(\omega)$ of the function as well as for the function itself, and $[X = x]$ is short for $[\omega: X(\omega) = x]$

A finite sum

$$(5.2) \qquad X = \sum_i x_i I_{A_i}$$

is a random variable if the A_i form a finite partition of Ω into $\mathscr{F}$-sets. Moreover, every simple random variable can be represented in the form (5.2): for the x_i take the range of X, and put $A_i = [X = x_i]$. But X may have other such representations because $x_i I_{A_i}$ can be replaced by $\sum_j x_i I_{A_{ij}}$ if the A_{ij} form a finite decomposition of A_i into $\mathscr{F}$-sets.

If $\mathscr{G}$ is a sub-σ-field of $\mathscr{F}$, a simple random variable X is *measurable* $\mathscr{G}$, or *measurable with respect to* $\mathscr{G}$, if $[X = x] \in \mathscr{G}$ for each x. A simple random variable is by definition always measurable $\mathscr{F}$. Since $[X \in H] = \bigcup [X = x]$, where the union extends over the finitely many x lying both in H and in the range of X, $[X \in H] \in \mathscr{G}$ for every $H \subset R^1$ if X is a simple random variable variable measurable $\mathscr{G}$.

The σ-field $\sigma(X)$ *generated by* X is the smallest σ-field with respect to which X is measurable; that is, $\sigma(X)$ is the intersection of all σ-fields with respect to which X is measurable. For a finite or infinite sequence $X_1, X_2, \ldots$ of simple random variables, $\sigma(X_1, X_2, \ldots)$ is the smallest σ-field with respect to which each X_i is measurable. It can be described explicitly in the finite case:

Theorem 5.1. *Let $X_1, \ldots, X_n$ be simple random variables.*

(i) *The σ-field $\sigma(X_1, \ldots, X_n)$ consists of the sets*

$$(5.3) \qquad [(X_1, \ldots, X_n) \in H] = [\omega: (X_1(\omega), \ldots, X_n(\omega)) \in H]$$

for $H \subset R^n$; H in this representation may be taken finite.

(ii) *A simple random variable Y is measurable $\sigma(X_1, \ldots, X_n)$ if and only if*

$$(5.4) \qquad Y = f(X_1, \ldots, X_n)$$

for some $f: R^n \to R^1$.

PROOF. Let $\mathscr{M}$ be the class of sets of the form (5.3). Sets of the form $[(X_1, \ldots, X_n) = (x_1, \ldots, x_n)] = \bigcap_{i=1}^n [X_i = x_i]$ must lie in $\sigma(X_1, \ldots, X_n)$; each set (5.3) is a finite union of sets of this form because $(X_1, \ldots, X_n)$, as a mapping from Ω to R^n, has finite range. Thus $\mathscr{M} \subset \sigma(X_1, \ldots, X_n)$.

On the other hand, $\mathscr{M}$ is a σ-field because $\Omega = [(X_1, \ldots, X_n) \in R^n]$, $[(X_1, \ldots, X_n) \in H]^c = [(X_1, \ldots, X_n) \in H^c]$, and $\bigcup_j [(X_1, \ldots, X_n) \in H_j] = [(X_1, \ldots, X_n) \in \bigcup_j H_j]$. But each X_i is measurable with respect to $\mathscr{M}$, because $[X_i = x]$ can be put in the form (5.3) by taking H to consist of those $(x_1, \ldots, x_n)$ in R^n for which $x_i = x$. It follows that $\sigma(X_1, \ldots, X_n)$ is contained

in $\mathcal{M}$ and therefore equals $\mathcal{M}$. As intersecting H with the range (finite) of $(X_1, \ldots, X_n)$ in R^n does not affect (5.3), H may be taken finite. This proves (i).

Assume that Y has the form (5.4)—that is, $Y(\omega) = f(X_1(\omega), \ldots, X_n(\omega))$ for every ω. Since $[Y = y]$ can be put in the form (5.3) by taking H to consist of those $x = (x_1, \ldots, x_n)$ for which $f(x) = y$, it follows that Y is measurable $\sigma(X_1, \ldots, X_n)$.

Now assume that Y is measurable $\sigma(X_1, \ldots, X_n)$. Let $y_1, \ldots, y_r$ be the distinct values Y assumes. By part (i), there exist sets $H_1, \ldots, H_r$ in R^n such that

$$\left[\omega : Y(\omega) = y_i \right] = \left[\omega : (X_1(\omega), \ldots, X_n(\omega)) \in H_i \right].$$

Take $f = \sum_{i=1}^{r} y_i I_{H_i}$. Although the H_i need not be disjoint, if H_i and H_j share a point of the form $(X_1(\omega), \ldots, X_n(\omega))$, then $Y(\omega) = y_i$ and $Y(\omega) = y_j$, which is impossible if $i \neq j$. Therefore each $(X_1(\omega), \ldots, X_n(\omega))$ lies in exactly one of the H_i, and it follows that $f(X_1(\omega), \ldots, X_n(\omega)) = Y(\omega)$. ∎

Since (5.4) implies that Y is measurable $\sigma(X_1, \ldots, X_n)$, it follows in particular that functions of simple random variables are again simple random variables. Thus X^2, e^{tX}, and so on are simple random variables along with X. Taking f to be $\sum_{i=1}^{n} x_i$, $\prod_{i=1}^{n} x_i$, or $\max_{i \leq n} x_i$ shows that sums, products, and maxima of simple random variables are simple random variables.

As explained on p. 57, a sub-σ-field corresponds to partial information about ω. From this point of view, $\sigma(X_1, \ldots, X_n)$ corresponds to a knowledge of the values $X_1(\omega), \ldots, X_n(\omega)$. These values suffice to determine the value $Y(\omega)$ if and only if (5.4) holds. The elements of the $\sigma(X_1, \ldots, X_n)$-partition (see (4.16)) are the sets $[X_1 = x_1, \ldots, X_n = x_n]$ for x_i in the range of X_i.

Example 5.1. For the dyadic digits $d_n(\omega)$ on the unit interval, d_3 is not measurable $\sigma(d_1, d_2)$; indeed, there exist ω' and ω'' such that $d_1(\omega') = d_1(\omega'')$ and $d_2(\omega') = d_2(\omega'')$ but $d_3(\omega') \neq d_3(\omega'')$, an impossibility if $d_3(\omega) = f(d_1(\omega), d_2(\omega))$ identically in ω. If such an f existed, one could unerringly predict the outcome $d_3(\omega)$ of the third toss from the outcomes $d_1(\omega)$ and $d_2(\omega)$ of the first two. ∎

Example 5.2. Let $s_n(\omega) = \sum_{k=1}^{n} r_k(\omega)$ be the partial sums of the Rademacher functions—see (1.14). By Theorem 5.1(ii) s_k is measurable $\sigma(r_1, \ldots, r_n)$ for $k \leq n$, and $r_k = s_k - s_{k-1}$ is measurable $\sigma(s_1, \ldots, s_n)$ for $k \leq n$. Thus $\sigma(r_1, \ldots, r_n) = \sigma(s_1, \ldots, s_n)$. In random-walk terms, the first n positions contain the same information as the first n distances moved. In gambling terms, to know the gambler's first n fortunes (relative to his initial fortune) is the same thing as to know his gains and losses on each of the first n plays. ∎

Example 5.3. An indicator I_A is measurable $\mathcal{I}$ if and only if A lies in $\mathcal{I}$. And $A \in \sigma(A_1, \ldots, A_n)$ if and only if $I_A = f(I_{A_1}, \ldots, I_{A_n})$ for some $f: R^n \to R^1$. ∎

Convergence of Random Variables

It is a basic problem, for given random variables X and $X_1, X_2, \ldots$ on a probability space $(\Omega, \mathcal{F}, P)$, to look for the probability of the event that $\lim_n X_n(\omega) = X(\omega)$. The normal number theorem is an example, one where the probability is 1. It is convenient to characterize the complementary event: $X_n(\omega)$ fails to converge to $X(\omega)$ if and only if there is some ϵ such that for no m does $|X_n(\omega) - X(\omega)|$ remain below ϵ for all n exceeding m—that is to say, if and only if, for some ϵ, $|X_n(\omega) - X(\omega)| \geq \epsilon$ holds for infinitely many values of n. Therefore,

$$(5.5) \qquad \left[\lim_n X_n = X \right]^c = \bigcup_\epsilon \left[|X_n - X| \geq \epsilon \text{ i.o.} \right],$$

where the union can be restricted to rational (positive) ϵ because the set in the union increases as ϵ decreases (compare (2.2)).

The event $[\lim_n X_n = X]$ therefore always lies in the basic σ-field $\mathcal{F}$, and it has probability 1 if and only if

$$(5.6) \qquad P\left[|X_n - X| \geq \epsilon \text{ i.o.} \right] = 0$$

for each ϵ (rational or not). The event in (5.6) is the limit superior of the events $[|X_n - X| \geq \epsilon]$, and it follows by Theorem 4.1 that (5.6) implies

$$(5.7) \qquad \lim_n P\left[|X_n - X| \geq \epsilon \right] = 0.$$

This leads to a definition: If (5.7) holds for each positive ϵ, then X_n is said to *converge to X in probability*, written $X_n \to_p X$.

These arguments prove two facts:

Theorem 5.2. (i) *There is convergence* $\lim_n X_n = X$ *with probability* 1 *if and only if* (5.6) *holds for each* ϵ.

(ii) *Convergence with probability* 1 *implies convergence in probability.*

Theorem 1.2, the normal number theorem, has to do with the convergence with probability 1 of $n^{-1}\sum_{i=1}^n d_i(\omega)$ to $\frac{1}{2}$. Theorem 1.1 has to do instead with the convergence in probability of the same sequence. By Theorem 5.2(ii), then, Theorem 1.1 is a consequence of Theorem 1.2 (see (1.30) and (1.31)). The converse is not true, however—convergence in probability does not imply convergence with probability 1:

Example 5.4. Take $X \equiv 0$ and $X_n = I_{A_n}$. Then $X_n \to_P X$ is equivalent to $P(A_n) \to 0$, and $[\lim_n X_n = X]^c = [A_n \text{ i.o.}]$. Any sequence $\{A_n\}$ such that $P(A_n) \to 0$ but $P[A_n \text{ i.o.}] > 0$ therefore gives a counterexample to the converse to Theorem 5.2(ii).

Consider the event $A_n = [\omega : l_n(\omega) \geq \log_2 n]$ in Example 4.15. Here, $P(A_n) \leq 1/n \to 0$, while $P[A_n \text{ i.o.}] = 1$ by (4.26), and so this is one counterexample. For an example more extreme and more transparent, define events in the unit interval in the following way. Define the first two by

$$A_1 = (0, \tfrac{1}{2}], \qquad A_2 = (\tfrac{1}{2}, 1].$$

Define the next four by

$$A_3 = (0, \tfrac{1}{4}], \qquad A_4 = (\tfrac{1}{4}, \tfrac{1}{2}], \qquad A_5 = (\tfrac{1}{2}, \tfrac{3}{4}], \qquad A_6 = (\tfrac{3}{4}, 1].$$

Define the next eight, $A_7, \ldots, A_{14}$, as the dyadic intervals of rank 3. And so on. Certainly, $P(A_n) \to 0$, and since each point ω is covered by one set in each successive block of length 2^k, the set $[A_n \text{ i.o.}]$ is all of $(0, 1]$. ∎

Independence

A sequence $X_1, X_2, \ldots$ (finite or infinite) of simple random variables is by definition *independent* if the classes $\sigma(X_1), \sigma(X_2), \ldots$ are independent in the sense of the preceding section. By Theorem 5.1(i), $\sigma(X_i)$ consists of the sets $[X_i \in H]$ for $H \subset R^1$. The condition for independence of $X_1, \ldots, X_n$ is therefore that

$$(5.8) \qquad P[X_1 \in H_1, \ldots, X_n \in H_n] = P[X_1 \in H_1] \cdots P[X_n \in H_n]$$

for linear sets $H_1, \ldots, H_n$. The definition (4.10) also requires that (5.8) hold if one or more of the $[X_i \in H_i]$ is suppressed; but taking H_i to be R^1 eliminates it from each side. For an infinite sequence $X_1, X_2, \ldots$, (5.8) must hold for each n. A special case of (5.8) is

$$(5.9) \qquad P[X_1 = x_1, \ldots, X_n = x_n] = P[X_1 = x_1] \cdots P[X_n = x_n].$$

On the other hand, summing (5.9) over $x_1 \in H_1, \ldots, x_n \in H_n$ gives (5.8). Thus the X_i are independent if and only if (5.9) holds for all $x_1, \ldots, x_n$.

Suppose that

$$(5.10) \qquad \begin{matrix} X_{11} & X_{12} & \cdots \\ X_{21} & X_{22} & \cdots \\ \vdots & \vdots & \end{matrix}$$

is an independent array of simple random variables. There may be finitely or

infinitely many rows, each row finite or infinite. If $\mathscr{A}_i$ consists of the finite intersections $\cap_j[X_{ij} \in H_j]$ with $H_j \subset R^1$, an application of Theorem 4.2 shows that the σ-fields $\sigma(X_{i1}, X_{i2}, \dots)$, $i = 1, 2, \dots$ are independent. As a consequence, $Y_1, Y_2, \dots$ are independent if Y_i is measurable $\sigma(X_{i1}, X_{i2}, \dots)$ for each i.

Example 5.5.   The dyadic digits $d_1(\omega), d_2(\omega), \dots$ on the unit interval are an independent sequence of random variables for which

$$(5.11) \qquad\qquad P[d_n = 0] = P[d_n = 1] = \tfrac{1}{2}.$$

It is because of (5 11) and independence that the d_n give a model for tossing a fair coin.

The sequence $(d_1(\omega), d_2(\omega), \dots)$ and the point ω determine one another. It can be imagined that ω is determined by the outcomes $d_n(\omega)$ of a sequence of tosses. It can also be imagined that ω is the result of drawing a point at random from the unit interval, and that ω determines the $d_n(\omega)$. In the second interpretation the $d_n(\omega)$ are all determined the instant ω is drawn, and so it should further be imagined that they are then revealed to the coin tosser or gambler one by one. For example, $\sigma(d_1, d_2)$ corresponds to knowing the outcomes of the first two tosses—to knowing not ω but only $d_1(\omega)$ and $d_2(\omega)$—and this does not help in predicting the value $d_3(\omega)$, because $\sigma(d_1, d_2)$ and $\sigma(d_3)$ are independent. See Example 5.1.   ∎

Example 5.6.   Every permutation can be written as a product of cycles. For example,

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 5 & 1 & 7 & 4 & 6 & 2 & 3 \end{pmatrix} = (1\,5\,6\,2)(3\,7)(4).$$

This permutation sends 1 to 5, 2 to 1, 3 to 7, and so on. The cyclic form on the right shows that 1 goes to 5, which goes to 6, which goes to 2, which goes back to 1; and so on. To standardize this cyclic representation, start the first cycle with 1 and each successive cycle with the smallest integer not yet encountered.

Let Ω consist of the $n!$ permutations of $1, 2, \dots, n$, all equally probable; $\mathscr{F}$ contains all subsets of Ω, and $P(A)$ is the fraction of points in A. Let $X_k(\omega)$ be 1 or 0 according as the element in the kth position in the cyclic representation of the permutation ω completes a cycle or not. Then $S(\omega) = \sum_{k=1}^n X_k(\omega)$ is the number of cycles in ω. In the example above, $n = 7$, $X_1 = X_2 = X_3 = X_5 = 0$, $X_4 = X_6 = X_7 = 1$, and $S = 3$. The following argument shows that $X_1, \dots, X_n$ are independent and $P[X_k = 1] = 1/(n - k + 1)$. This will lead later on to results on $P[S \in H]$.

The idea is this: $X_1(\omega) = 1$ if and only if the random permutation ω sends 1 to itself, the probability of which is $1/n$. If it happens that $X_1(\omega) = 1$—that ω fixes 1—then the image of 2 is one of $2, \ldots, n$, and $X_2(\omega) = 1$ if and only if this image is in fact 2; the conditional probability of this is $1/(n-1)$. If $X_1(\omega) = 0$, on the other hand, then ω sends 1 to some $i \neq 1$, so that the image of i is one of $1, \ldots, i-1, i+1, \ldots, n$, and $X_2(\omega) = 1$ if and only if this image is in fact 1; the conditional probability of this is again $1/(n-1)$. This argument generalizes.

But the details are fussy. Let $Y_1(\omega), \ldots, Y_n(\omega)$ be the integers in the successive positions in the cyclic representation of ω. Fix k, and let A_t be the set where $(X_1, \ldots, X_{k-1}, Y_1, \ldots, Y_k)$ assumes a specific vector of values $v = (x_1, \ldots, x_{k-1}, y_1, \ldots, y_k)$. The A_t form a partition $\mathscr{A}$ of Ω, and if $P[X_k = 1 | A_t] = 1/(n-k+1)$ for each v, then by Example 4.7, $P[X_k = 1] = 1/(n-k+1)$ and X_k is independent of $\sigma(\mathscr{A})$ and hence of the smaller σ-field $\sigma(X_1, \ldots, X_{k-1})$. It will follow by induction that $X_1, \ldots, X_n$ are independent.

Let j be the position of the rightmost 1 among $x_1, \ldots, x_{k-1}$ ($j = 0$ if there are none). Then ω lies in A_t if and only if it permutes $y_1, \ldots, y_j$ among themselves (in a way specified by the values $x_1, \ldots, x_{j-1}, x_j = 1, y_1, \ldots, y_j$) and sends each of $y_{j+1}, \ldots, y_{k-1}$ to the y just to its right. Thus A_t contains $(n-k+1)!$ sample points. And $X_k(\omega) = 1$ if and only if ω also sends y_k to y_{j+1}. Thus $A_t \cap [X_k = 1]$ contains $(n-k)!$ sample points, and so the conditional probability of $X_k = 1$ is $1/(n-k+1)$. ∎

Existence of Independent Sequences

The *distribution* of a simple random variable X is the probability measure μ defined for all subsets A of the line by

$$(5.12) \qquad \mu(A) = P[X \in A].$$

This does define a probability measure. It is discrete in the sense of Example 2.9: If $x_1, \ldots, x_l$ are the distinct points of the range of X, then μ has mass $p_i = P[X = x_i] = \mu\{x_i\}$ at x_i, and $\mu(A) = \sum p_i$, the sum extending over those i for which $x_i \in A$. As $\mu(A) = 1$ if A is the range of X, not only is μ discrete, it has finite support.

Theorem 5.3. *Let $\{\mu_n\}$ be a sequence of probability measures on the class of all subsets of the line, each having finite support. There exists on some probability space $(\Omega, \mathscr{F}, P)$ an independent sequence $\{X_n\}$ of simple random variables such that X_n has distribution μ_n.*

What matters here is that there are finitely or countably many distributions μ_n. They need not be indexed by the integers; any countable index set will do.

PROOF. The probability space will be the unit interval. To understand the construction, consider first the case in which each μ_n concentrates its mass on the two points 0 and 1. Put $p_n = \mu_n\{0\}$ and $q_n = 1 - p_n = \mu_n\{1\}$. Split $(0, 1]$ into two intervals I_0 and I_1 of lengths p_1 and q_1. Define $X_1(\omega) = 0$ for $\omega \in I_0$ and $X_1(\omega) = 1$ for $\omega \in I_1$. If P is Lebesgue measure, then clearly $P[X_1 = 0] = p_1$ and $P[X_1 = 1] = q_1$, so that X_1 has distribution μ_1.

$X_1 = 0$	$X_1 = 1$
p_1	q_1

| $X_1 = 0$ | $X_1 = 0$ | $X_1 = 1$ | $X_1 = 1$ |
$X_2 = 0$	$X_2 = 1$	$X_2 = 0$	$X_2 = 1$
$p_1 p_2$	$p_1 q_2$	$q_1 p_2$	$q_1 q_2$
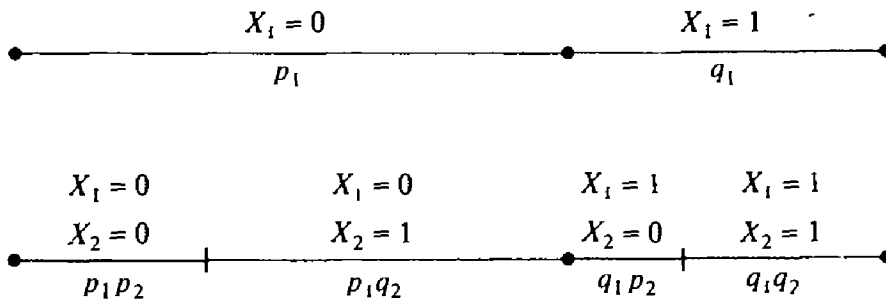
Now split I_0 into two intervals I_{00} and I_{01} of lengths $p_1 p_2$ and $p_1 q_2$, and split I_1 into two intervals I_{10} and I_{11} of lengths $q_1 p_2$ and $q_1 q_2$. Define $X_2(\omega) = 0$ for $\omega \in I_{00} \cup I_{10}$ and $X_2(\omega) = 1$ for $\omega \in I_{01} \cup I_{11}$. As the diagram makes clear, $P[X_1 = 0, \ X_2 = 0] = p_1 p_2$, and similarly for the other three possibilities. It follows that X_1 and X_2 are independent and X_2 has distribution μ_2. Now X_3 is constructed by splitting each of $I_{00}, I_{01}, I_{10}, I_{11}$ in the proportions p_3 and q_3. And so on.

If $p_n = q_n = \frac{1}{2}$ for all n, then the successive decompositions here are the decompositions of $(0, 1]$ into dyadic intervals, and $X_n(\omega) = d_n(\omega)$.

The argument for the general case is not very different. Let $x_{n1}, \ldots, x_{nl_n}$ be the distinct points on which μ_n concentrates its mass, and put $p_{ni} = \mu_n\{x_{ni}\}$ for $1 \leq i \leq l_n$.

Decompose[†] $(0, 1]$ into l_1 subintervals $I_1^{(1)}, \ldots, I_{l_1}^{(1)}$ of respective lengths $p_{11}, \ldots, p_{1l_1}$. Define X_1 by setting $X_1(\omega) = x_{1i}$ for $\omega \in I_i^{(1)}$, $1 \leq i \leq l_1$. Then (P is Lebesgue measure) $P[\omega: X_1(\omega) = x_{1i}] = P(I_i^{(1)}) = p_{1i}$, $1 \leq i \leq l_1$. Thus X_1 is a simple random variable with distribution μ_1.

Next decompose each $I_i^{(1)}$ into l_2 subintervals $I_{i1}^{(2)}, \ldots, I_{il_2}^{(2)}$ of respective lengths $p_{1i}p_{21}, \ldots, p_{1i}p_{2l_2}$. Define $X_2(\omega) = x_{2j}$ for $\omega \in \bigcup_{i=1}^{l_1} I_{ij}^{(2)}$, $1 \leq j \leq l_2$. Then $P[\omega: X_1(\omega) = x_{1i}, X_2(\omega) = x_{2j}] = P(I_{ij}^{(2)}) = p_{1i}p_{2j}$. Adding out i shows that $P[\omega: X_2(\omega) = x_{2j}] = p_{2j}$, as required. Hence $P[X_1 = x_{1i}, X_2 = x_{2j}] = p_{1i}p_{2j} = P[X_1 = x_{1i}]P[X_2 = x_{2j}]$, and X_1 and X_2 are independent.

The construction proceeds inductively. Suppose that $(0, 1]$ has been decomposed into $l_1 \cdots l_n$ intervals

$$(5.13) \qquad I_{i_1 \cdots i_n}^{(n)}, \qquad 1 \leq i_1 \leq l_1, \ldots, 1 \leq i_n \leq l_n,$$

[†] If $b - a = \delta_1 + \cdots + \delta_l$ and $\delta_i \geq 0$, then $I_i = (a + \sum_{j<i}\delta_j, \ a + \sum_{j \leq i}\delta_j]$ decomposes $(a, b]$ into subintervals $I_1, \ldots, I_l$ with lengths of δ_i. Of course, I_i is empty if $\delta_i = 0$.

of lengths

(5.14) $P\left(I_{i_1}^{(n)} {}_{i_n}\right) = p_{1,i_1} \cdots p_{n,i_n}.$

Decompose $I_{i_1}^{(n)} {}_{i_n}$ into l_{n+1} subintervals $I_{i_1}^{(n+1)} {}_{i_n 1}, \ldots, I_{i_1}^{(n+1)} {}_{i_n l_{n+1}}$ of respective lengths $P(I_{i_1}^{(n)} {}_{i_n}) p_{n+1,1}, \ldots, P(I_{i_1}^{(n)} {}_{i_n}) p_{n+1, l_{n+1}}$. These are the intervals of the next decomposition. This construction gives a sequence of decompositions (5.13) of $(0, 1]$ into subintervals; each decomposition satisfies (5.14), and each refines the preceding one. If μ_n is given for $1 \le n \le N$, the procedure terminates after N steps; for an infinite sequence it does not terminate at all.

For $1 \le i \le l_n$, put $X_n(\omega) = x_{ni}$ if $\omega \in \bigcup_{i_1} {}_{i_{n-1}} I_{i_1}^{(n)} {}_{i_{n-1}i}$. Since each decomposition (5.13) refines the preceding, $X_k(\omega) = x_{ki_k}$ for $\omega \in I_{i_1}^{(n)} {}_{i_k} {}_{i_n}$. Therefore, each element of (5.13) is contained in the element with the same label $i_1 \ldots i_n$ in the decomposition

$$A_{i_1 \ \ i_n} = \left[\omega : X_1(\omega) = x_{1i_1}, \ldots, X_n(\omega) = x_{ni_n}\right]. \quad 1 \le i_1 \le l_1, \ldots, \quad 1 \le i_n \le l_n.$$

The two decompositions thus coincide, and it follows by (5.14) that $P\{X_1 = x_{1i_1}, \ldots, X_n = x_{ni_n}\} = p_{1,i_1} \cdots p_{n,i_n}$. Adding out the indices $i_1, \ldots, i_{n-1}$ shows that X_n has distribution μ_n and hence that $X_1, \ldots, X_n$ are independent. But n was arbitrary.   ∎

In the case where the μ_n are all the same, there is an alternative construction based on probabilities in sequence space. Let S be the support (finite) common to the μ_n, and let p_u, $u \in S$, be the probabilities common to the μ_n. In sequence space S^∞, define product measure P on the class $\mathscr{C}_0$ of cylinders by (2.21). By Theorem 2.3, P is countably additive on $\mathscr{C}_0$, and by Theorem 3.1 it extends to $\mathscr{C} = \sigma(\mathscr{C}_0)$. The coordinate functions $z_k(\cdot)$ are random variables on the probability space $(S^\infty, \mathscr{C}, P)$; take these as the X_k. Then (2.22) translates into $P[X_1 = u_1, \ldots, X_n = u_n] = p_{u_1} \cdots p_{u_n}$, which is just what Theorem 5.3 requires in this special case.

Probability theorems such as those in the next sections concern independent sequences $\{X_n\}$ with specified distributions or with distributions having specified properties, and because of Theorem 5.3 these theorems are true not merely in the vacuous sense that their hypotheses are never fulfilled. Similar but more complicated existence theorems will come later. For most purposes the probability space on which the X_n are defined is largely irrelevant. Every independent sequence $\{X_n\}$ satisfying $P\{X_n = 1\} = p$ and $P[X_n = 0] = 1 - p$ is a model for Bernoulli trials, for example, and for an event like $\bigcup_{n=1}^\infty [\sum_{k=1}^n X_k > \alpha n]$, expressed in terms of the X_n alone, the calculation of its probability proceeds in the same way whatever the underlying space $(\Omega, \mathscr{F}, P)$ may be.

It is, of course, an advantage that such results apply not just to some canonical sequence $\{X_n\}$ (such as the one constructed in the proof above) but to every sequence with the appropriate distributions. In some applications of probability within mathematics itself, such as the arithmetic applications of run theory in the preceding section, the underlying Ω does play a role.

Expected Value

A simple random variable in the form (5.2) is assigned *expected value* or *mean value*

$$(5.15) \qquad E[X] = E\left[\sum_i x_i I_{A_i}\right] = \sum_i x_i P(A_i).$$

There is the alternative form

$$(5.16) \qquad E[X] = \sum_x x P[X = x],$$

the sum extending over the range of X; indeed, (5.15) and (5.16) both coincide with $\sum_x \sum_{i:\, x_i = x} x_i P(A_i)$. By (5.16) the definition (5.15) is consistent: different representations (5.2) give the same value to (5.15). From (5.16) it also follows that $E[X]$ depends only on the distribution of X; hence $E[X] = E[Y]$ if $P[X = Y] = 1$.

If X is a simple random variable on the unit interval and if the A_i in (5.2) happen to be subintervals, then (5.15) coincides with the Riemann integral as given by (1.6). More general notions of integral and expected value will be studied later. Simple random variables are easy to work with because the theory of their expected values is transparent and free of technical complications.

As a special case of (5.15) and (5.16),

$$(5.17) \qquad E[I_A] = P(A).$$

As another special case, if a constant α is identified with the random variable $X(\omega) \equiv \alpha$, then

$$(5.18) \qquad E[\alpha] = \alpha.$$

From (5.2) follows $f(X) = \sum_i f(x_i) I_{A_i}$, and hence

$$(5.19) \qquad E[f(X)] = \sum_i f(x_i) P(A_i) = \sum_x f(x) P[X = x],$$

the last sum extending over the range of X. For example, the kth *moment* $E[X^k]$ of X is defined by $E[X^k] = \sum_y y P[X^k = y]$, where y varies over the

range of X^k, but it is usually simpler to compute it by $E[X^k] = \sum_x x^k P[X = x]$, where x varies over the range of X.

If

$$(5.20) \qquad X = \sum_i x_i I_{A_i}, \qquad Y = \sum_j y_j I_{B_j}$$

are simple random variables, then $\alpha X + \beta Y = \sum_{ij} (\alpha x_i + \beta y_j) I_{A_i \cap B_j}$ has expected value $\sum_{ij}(\alpha x_i + \beta y_j) P(A_i \cap B_j) = \alpha \sum_i x_i P(A_i) + \beta \sum_j y_j P(B_j)$. Expected value is therefore *linear*:

$$(5.21) \qquad E[\alpha X + \beta Y] = \alpha E[X] + \beta E[Y].$$

If $X(\omega) \le Y(\omega)$ for all ω, then $x_i \le y_j$ if $A_i \cap B_j$ is nonempty, and hence $\sum_{ij} x_i P(A_i \cap B_j) \le \sum_{ij} y_j P(A_i \cap B_j)$. Expected value therefore *preserves order*:

$$(5.22) \qquad E[X] \le E[Y] \qquad \text{if } X \le Y.$$

(It is enough that $X \le Y$ on a set of probability 1.) Two applications of (5.22) give $E[-|X|] \le E[X] \le E[|X|]$, so that by linearity,

$$(5.23) \qquad |E[X]| \le E[|X|].$$

And more generally,

$$(5.24) \qquad |E[X - Y]| \le E[|X - Y|].$$

The relations (5.17) through (5.24) will be used repeatedly, and so will the following theorem on expected values and limits. If there is a finite K such that $|X_n(\omega)| \le K$ for all ω and n, the X_n are *uniformly bounded*.

Theorem 5.4. *If $\{X_n\}$ is uniformly bounded, and if $X = \lim_n X_n$ with probability* 1, *then* $E[X] = \lim_n E[X_n]$.

PROOF. By Theorem 5.2(ii), convergence with probability 1 implies convergence in probability: $X_n \to_P X$. And in fact the latter suffices for the present proof. Increase K so that it bounds $|X|$ (which has finite range) as well as all the $|X_n|$; then $|X - X_n| \le 2K$. If $A = [|X - X_n| \ge \epsilon]$, then

$$|X(\omega) - X_n(\omega)| \le 2K I_A(\omega) + \epsilon I_{A^c}(\omega) \le 2K I_A(\omega) + \epsilon$$

for all ω. By (5.17), (5.18), (5.21), and (5.22),

$$E[|X - X_n|] \le 2K P[|X - X_n| \ge \epsilon] + \epsilon.$$

But since $X_n \to_P X$, the first term on the right goes to 0, and since ϵ is arbitrary, $E[|X - X_n|] \to 0$. Now apply (5.24). ∎

Theorems of this kind are of constant use in probability and analysis. For the general version, Lebesgue's dominated convergence theorem, see Section 16.

Example 5.7. On the unit interval, take $X(\omega)$ identically 0, and take $X_n(\omega)$ to be n^2 if $0 < \omega \le n^{-1}$ and 0 if $n^{-1} < \omega \le 1$. Then $X_n(\omega) \to X(\omega)$ for every ω, although $E[X_n] = n$ does not converge to $E[X] = 0$. Thus theorem 5.4 fails without some hypothesis such as that of uniform boundedness. See also Example 7.7. ∎

An extension of (5.21) is an immediate consequence of Theorem 5.4:

Corollary. *If* $X = \sum_n X_n$ *on an $\mathscr{F}$-set of probability* 1, *and if the partial sums of* $\sum_n X_n$ *are uniformly bounded, then* $E[X] = \sum_n E[X_n]$.

Expected values for independent random variables satisfy the familiar product law. For X and Y as in (5.20), $XY = \sum_{ij} x_i y_j I_{A_i \cap B_j}$. If the x_i are distinct and the y_j are distinct, then $A_i = [X = x_i]$ and $B_j = [Y = y_j]$; for independent X and Y, $P(A_i \cap B_j) = P(A_i)P(B_j)$ by (5.9), and so $E[XY] = \sum_{ij} x_i y_j P(A_i)P(B_j) = E[X]E[Y]$. If X, Y, Z are independent, then XY and Z are independent by the argument involving (5.10), so that $E[XYZ] = E[XY]E[Z] = E[X]E[Y]E[Z]$. This obviously extends:

$$(5.25) \qquad E[X_1 \cdots X_n] = E[X_1] \cdots E[X_n]$$

if $X_1, \ldots, X_n$ are independent.

Various concepts from discrete probability carry over to simple random variables. If $E[X] = m$, the *variance* of X is

$$(5.26) \qquad \mathrm{Var}[X] = E\left[(X - m)^2\right] = E[X^2] - m^2;$$

the left-hand equality is a definition, the right-hand one a consequence of expanding the square. Since $\alpha X + \beta$ has mean $\alpha m + \beta$, its variance is $E[((\alpha X + \beta) - (\alpha m + \beta))^2] = E[\alpha^2(X - m)^2]$:

$$(5.27) \qquad \mathrm{Var}[\alpha X + \beta] = \alpha^2 \mathrm{Var}[X].$$

If $X_1, \ldots, X_n$ have means $m_1, \ldots, m_n$, then $S = \sum_{i=1}^{n} X_i$ has mean $m = \sum_{i=1}^{n} m_i$, and $E[(S - m)^2] = E[(\sum_{i=1}^{n}(X_i - m_i))^2] = \sum_{i=1}^{n} E[(X_i - m_i)^2] + 2\sum_{1 \le i < j \le n} E[(X_i - m_i)(X_j - m_j)]$. If the X_i are independent, then so are the $X_i - m_i$, and by (5.25) the last sum vanishes. This gives the familiar formula

for the variance of a sum of independent random variables:

$$(5.28) \qquad \operatorname{Var}\left[\sum_{i=1}^{n} X_i \right] = \sum_{i=1}^{n} \operatorname{Var}[X_i].$$

Suppose that X is nonnegative; order its range: $0 \le x_1 < x_2 < \cdots < x_k$. Then

$$E[X] = \sum_{i=1}^{k} x_i P[X = x_i]$$

$$= \sum_{i=1}^{k-1} x_i (P[X \ge x_i] - P[X \ge x_{i+1}]) + x_k P[X \ge x_k]$$

$$= x_1 P[X \ge x_1] + \sum_{i=2}^{k} (x_i - x_{i-1}) P[X \ge x_i].$$

Since $P[X \ge x] = P[X \ge x_1]$ for $0 \le x \le x_1$ and $P[X \ge x] = P[X \ge x_i]$ for $x_{i-1} < x \le x_i$, it is possible to write the final sum as the Riemann integral of a step function:

$$(5.29) \qquad E[X] = \int_0^\infty P[X \ge x]\, dx.$$

This holds if X is nonnegative. Since $P[X \ge x] = 0$ for $x > x_k$, the range of integration is really finite.

There is for (5.29) a simple geometric argument involving the "area over the curve." If $p_i = P[X = x_i]$, the area of the shaded region in the figure is the sum $p_1 x_1 + \cdots + p_k x_k = E[X]$ of the areas of the horizontal strips; it is also the integral of the height $P[X \ge x]$ of the region.
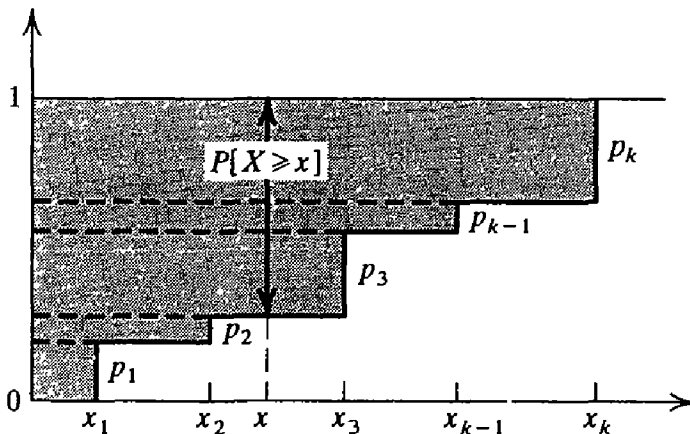
Inequalities

There are for expected values several standard inequalities that will be needed. If X is nonnegative, then for positive α (sum over the range of X) $E[X] = \sum_x xP[X = x] \geq \sum_{x\ x \geq \alpha} xP[X = x] \geq \alpha \sum_{x\ x \geq \alpha} P[X = x]$. Therefore,

$$(5.30) \qquad P[X \geq \alpha] \leq \frac{1}{\alpha} E[X]$$

if X is nonnegative and α positive. A special case of this is (1.20). Applied to $|X|^k$, (5.30) gives *Markov's inequality*,

$$(5.31) \qquad P[|X| \geq \alpha] \leq \frac{1}{\alpha^k} E[|X|^k],$$

valid for positive α. If $k = 2$ and $m = E[X]$ is subtracted from X, this becomes the *Chebyshev* (or Chebyshev–Bienaymé) *inequality*:

$$(5.32) \qquad P[|X - m| \geq \alpha] \leq \frac{1}{\alpha^2} \operatorname{Var}[X].$$

A function φ on an interval is *convex* [A32] if $\varphi(px + (1 - p)y) \leq p\varphi(x) + (1 - p)\varphi(y)$ for $0 \leq p \leq 1$ and x and y in the interval. A sufficient condition for this is that φ have a nonnegative second derivative. It follows by induction that $\varphi(\sum_{i=1}^l p_i x_i) \leq \sum_{i=1}^l p_i \varphi(x_i)$ if the p_i are nonnegative and add to 1 and the x_i are in the domain of φ. If X assumes the value x_i with probability p_i, this becomes *Jensen's inequality*,

$$(5.33) \qquad \varphi(E[X]) \leq E[\varphi(X)],$$

valid if φ is convex on an interval containing the range of X.

Suppose that

$$(5.34) \qquad \frac{1}{p} + \frac{1}{q} = 1, \qquad p > 1, \quad q > 1.$$

Hölder's inequality is

$$(5.35) \qquad E[|XY|] \leq E^{1/p}[|X|^p] \cdot E^{1/q}[|Y|^q].$$

If, say, the first factor on the right vanishes, then $X = 0$ with probability 1, hence $XY = 0$ with probability 1, and hence the left side vanishes also. Assume then that the right side of (5.35) is positive. If a and b are positive, there exist s and t such that $a = e^{p^{-1}s}$ and $b = e^{q^{-1}t}$. Since e^x is convex,

$$e^{p^{-1}s + q^{-1}t} \le p^{-1}e^s + q^{-1}e^t, \text{ or}$$

$$ab \le \frac{a^p}{p} + \frac{b^q}{q}.$$

This obviously holds for nonnegative as well as for positive a and b. Let u and v be the two factors on the right in (5.35). For each ω,

$$\left| \frac{X(\omega)Y(\omega)}{uv} \right| \le \frac{1}{p} \left| \frac{X(\omega)}{u} \right|^p + \frac{1}{q} \left| \frac{Y(\omega)}{v} \right|^q$$

Taking expected values and applying (5.34) leads to (5.35).

If $p = q = 2$, Hölder's inequality becomes *Schwarz's inequality*:

(5.36) $$E[|XY|] \le E^{1/2}[X^2] \cdot E^{1/2}[Y^2].$$

Suppose that $0 < \alpha < \beta$. In (5.35) take $p = \beta/\alpha$, $q = \beta/(\beta - \alpha)$, and $Y(\omega) = 1$, and replace X by $|X|^\alpha$. The result is *Lyapounov's inequality*,

(5.37) $$E^{1/\alpha}[|X|^\alpha] \le E^{1/\beta}[|X|^\beta], \qquad 0 < \alpha \le \beta.$$

PROBLEMS

5.1. (a) Show that X is measurable with respect to the σ-field $\mathscr{G}$ if and only if $\sigma(X) \subset \mathscr{G}$. Show that X is measurable $\sigma(Y)$ if and only if $\sigma(X) \subset \sigma(Y)$.

(b) Show that, if $\mathscr{G} = \{\varnothing, \Omega\}$, then X is measurable $\mathscr{G}$ if and only if X is constant.

(c) Suppose that $P(A)$ is 0 or 1 for every A in $\mathscr{G}$. This holds, for example, if $\mathscr{G}$ is the tail field of an independent sequence (Theorem 4.5), or if $\mathscr{G}$ consists of the countable and cocountable sets on the unit interval with Lebesgue measure. Show that if X is measurable $\mathscr{G}$, then $P[X = c] = 1$ for some constant c.

5.2. 2.19↑   Show that the unit interval can be replaced by any nonatomic probability measure space in the proof of Theorem 5.3.

5.3. Show that $m = E[X]$ minimizes $E[(X - m)^2]$.

5.4. Suppose that X assumes the values $m - \alpha, m, m + \alpha$ with probabilities $p, 1 - 2p, p$, and show that there is equality in (5.32). Thus Chebyshev's inequality cannot be improved without special assumptions on X.

5.5. Suppose that X has mean m and variance σ^2.

(a) Prove *Cantelli's inequality*

$$P[X - m \ge \alpha] \le \frac{\sigma^2}{\sigma^2 + \alpha^2}, \qquad \alpha \ge 0.$$

(b) Show that $P[|X - m| \geq \alpha] \leq 2\sigma^2/(\sigma^2 + \alpha^2)$. When is this better than Chebyshev's inequality?

(c) By considering a random variable assuming two values, show that Cantelli's inequality is sharp.

5.6. The polynomial $E[(t|X| + |Y|)^2]$ in t has at most one real zero. Deduce Schwarz's inequality once more.

5.7. (a) Write (5.37) in the form $E^{\beta/q}[|X|^\alpha] \leq E[|X|^\alpha)^{\beta/\alpha}]$ and deduce it directly from Jensen's inequality.

(b) Prove that $E[1/X^p] \geq 1/E^p[X]$ for $p > 0$ and X a positive random variable.

5.8. (a) Let f be a convex real function on a convex set C in the plane. Suppose that $(X(\omega), Y(\omega)) \in C$ for all ω and prove a two-dimensional Jensen's inequality:

$$(5.38) \qquad f(E[X], E[Y]) \leq E[f(X, Y)].$$

(b) Show that f is convex if it has continuous second derivatives that satisfy

$$(5.39) \qquad f_{11} \geq 0, \qquad f_{22} \geq 0, \qquad f_{11} f_{22} \geq f_{12}^2.$$

5.9. ↑ Hölder's inequality is equivalent to $E[X^{1/p} Y^{1/q}] \leq E^{1/p}[X] \cdot E^{1/q}[Y]$ $(p^{-1} + q^{-1} = 1)$, where X and Y are nonnegative random variables. Derive this from (5.38).

5.10. ↑ *Minkowski's inequality* is

$$(5.40) \qquad E^{1/p}[|X + Y|^p] \leq E^{1/p}[|X|^p] + E^{1/p}[|Y|^p],$$

valid for $p \geq 1$. It is enough to prove that $E[(X^{1/p} + Y^{1/p})^p] \leq (E^{1/p}[X] + E^{1/p}[Y])^p$ for nonnegative X and Y. Use (5.38).

5.11. For events $A_1, A_2, \ldots$, not necessarily independent, let $N_n = \sum_{k=1}^n I_{A_k}$ be the number to occur among the first n. Let

$$(5.41) \quad \alpha_n = \frac{1}{n} \sum_{k=1}^n P(A_k), \qquad \beta_n = \frac{2}{n(n-1)} \sum_{1 \leq j < k \leq n} P(A_j \cap A_k).$$

Show that

$$(5.42) \qquad E[n^{-1} N_n] = \alpha_n, \qquad \text{Var}[n^{-1} N_n] = \beta_n - \alpha_n^2 + \frac{\alpha_n - \beta_n}{n}.$$

Thus $\text{Var}[n^{-1} N_n] \to 0$ if and only if $\beta_n - \alpha_n^2 \to 0$, which holds if the A_n are independent and $P(A_n) = p$ (Bernoulli trials), because then $\alpha_n = p$ and $\beta_n = p^2 = \alpha_n^2$.

5.12. Show that, if X has nonnegative integers as values, then $E[X] = \sum_{n=1}^{\infty} P[X \geq n]$.

5.13. Let $I_i = I_{A_i}$ be the indicators of n events having union A. Let $S_k = \sum I_{i_1} \cdots I_{i_k}$, where the summation extends over all k-tuples satisfying $1 \leq i_1 < \cdots < i_k \leq n$. Then $s_k = E[S_k]$ are the terms in the inclusion–exclusion formula $P(A) = s_1 - s_2 + \cdots \pm s_n$. Deduce the inclusion–exclusion formula from $I_A = S_1 - S_2 + \cdots \pm S_n$. Prove the latter formula by expanding the product $\prod_{i=1}^{n}(1 - I_i)$

5.14. Let $f_n(x)$ be $n^2 x$ or $2n - n^2 x$ or 0 according as $0 \leq x \leq n^{-1}$ or $n^{-1} \leq x \leq 2n^{-1}$ or $2n^{-1} \leq x \leq 1$. This gives a standard example of a sequence of continuous functions that converges to 0 but not uniformly. Note that $\int_0^1 f_n(x)\,dx$ does not converge to 0; relate to Example 5.7.

5.15. By Theorem 5.3, for any prescribed sequence of probabilities p_n, there exists (on some space) an independent sequence of events A_n satisfying $P(A_n) = p_n$. Show that if $p_n \to 0$ but $\sum p_n = \infty$, this gives a counterexample (like Example 5.4) to the converse of Theorem 5.2(ii).

5.16. ↑ Suppose that $0 \leq p_n \leq 1$ and put $\alpha_n = \min\{p_n, 1 - p_n\}$. Show that, if $\sum \alpha_n$ converges, then on some discrete probability space there exist independent events A_n satisfying $P(A_n) = p_n$. Compare Problem 1.1(b).

5.17. (a) Suppose that $X_n \to_P X$ and that f is continuous. Show that $f(X_n) \to_P f(X)$.
(b) Show that $E[|X - X_n|] \to 0$ implies $X_n \to_P X$. Show that the converse is false.

5.18. 2.20 ↑ The proof given for Theorem 5.3 for the special case where the μ_n are all the same can be extended to cover the general case: use Problem 2.20.

5.19. 2.18 ↑ For integers m and primes p, let $\alpha_p(m)$ be the exact power of p in the prime factorization of m: $m = \prod_p p^{\alpha_p(m)}$. Let $\delta_p(m)$ be 1 or 0 as p divides m or not. Under each P_n (see (2.34)) the α_p and δ_p are random variables. Show that for distinct primes $p_1, \ldots, p_u$,

$$(5.43) \qquad P_n\big[\alpha_{p_i} \geq k_i, i \leq u\big] = \frac{1}{n}\left\lfloor \frac{n}{p_1^{k_1} \cdots p_u^{k_u}} \right\rfloor \to \frac{1}{p_1^{k_1} \cdots p_u^{k_u}}$$

and

$$(5.44) \qquad P_n\big[\alpha_{p_i} = k_i, i \leq u\big] \to \prod_{i=1}^{u}\left(\frac{1}{p_i^{k_i}} - \frac{1}{p_i^{k_i+1}}\right).$$

Similarly,

$$(5.45) \qquad P_n\big[\delta_{p_i} = 1, i \leq u\big] = \frac{1}{n}\left\lfloor \frac{n}{p_1 \cdots p_u} \right\rfloor \to \frac{1}{p_1 \cdots p_u}.$$

According to (5.44), the α_p are for large n approximately independent under P_n, and according to (5.45), the same is true of the δ_p.

For a function f of positive integers, let

$$(5.46) \qquad E_n[f] = \frac{1}{n} \sum_{m=1}^{n} f(m)$$

be its expected value under the probability measure P_n. Show that

$$(5.47) \qquad E_n[\alpha_p] = \sum_{k=1}^{\infty} \frac{1}{n} \left\lfloor \frac{n}{p^k} \right\rfloor \to \frac{1}{p-1};$$

this says roughly that $(p-1)^{-1}$ is the average power of p in the factorization of large integers.

5.20. ↑ (a) From Stirling's formula, deduce

$$(5.48) \qquad E_n[\log] = \log n + O(1).$$

From this, the inequality $E_n[\alpha_p] \le 2/p$, and the relation $\log m = \sum_p \alpha_p(m) \log p$, conclude that $\sum_p p^{-1} \log p$ diverges and that there are infinitely many primes.
(b) Let $\log^* m = \sum_p \delta_p(m) \log p$. Show that

$$(5.49) \qquad E_n[\log^*] = \sum_p \frac{1}{n} \left\lfloor \frac{n}{p} \right\rfloor \log p = \log n + O(1).$$

(c) Show that $\lfloor 2n/p \rfloor - 2\lfloor n/p \rfloor$ is always nonnegative and equals 1 in the range $n < p \le 2n$. Deduce $E_{2n}[\log^*] - E_n[\log^*] = O(1)$ and conclude that

$$(5.50) \qquad \sum_{p \le r} \log p = O(x).$$

Use this to estimate the error removing the integral-part brackets introduces into (5.49), and show that

$$(5.51) \qquad \sum_{p \le x} p^{-1} \log p = \log x + O(1).$$

(d) Restrict the range of summation in (5.51) to $\theta x < p \le x$ for an appropriate θ, and conclude that

$$(5.52) \qquad \sum_{p \le x} \log p \asymp x,$$

in the sense that the ratio of the two sides is bounded away from 0 and ∞.
(e) Use (5.52) and truncation arguments to prove for the number $\pi(x)$ of primes not exceeding x that

$$(5.53) \qquad \pi(x) \asymp \frac{x}{\log x}.$$

(By the prime number theorem the ratio of the two sides in fact goes to 1.)
Conclude that the rth prime p_r satisfies $p_r \asymp r \log r$ and that

(5.54)
$$\sum_p \frac{1}{p} = \infty.$$

SECTION 6. THE LAW OF LARGE NUMBERS

The Strong Law

Let $X_1, X_2, \ldots$ be a sequence of simple random variables on some probability space $(\Omega, \mathscr{F}, P)$. They are *identically distributed* if their distributions (in the sense of (5.12)) are all the same. Define $S_n = X_1 + \cdots + X_n$. The *strong law of large numbers*:

Theorem 6.1. *If the X_n are independent and identically distributed and* $E[X_n] = m$, *then*

(6.1)
$$P\left[\lim_n n^{-1} S_n = m \right] = 1.$$

PROOF. The conclusion is that $n^{-1}S_n - m = n^{-1}\sum_{i=1}^n (X_i - m) \to 0$ with probability 1. Replacing X_i by $X_i - m$ shows that there is no loss of generality in assuming that $m = 0$. The set in question does lie in $\mathscr{F}$ (see (5.5)), and by Theorem 5.2(i), it is enough to show that $P[|n^{-1}S_n| \ge \epsilon \text{ i.o.}] = 0$ for each ϵ.

Let $E[X_i^2] = \sigma^2$ and $E[X_i^4] = \xi^4$. The proof is like that for Theorem 1.2. First (see (1.26)), $E[S_n^4] = \sum E[X_\alpha X_\beta X_\gamma X_\delta]$, the four indices ranging independently from 1 to n. Since $E[X_i] = 0$, it follows by the product rule (5.25) for independent random variables that the summand vanishes if there is one index different from the three others. This leaves terms of the form $E[X_i^4] = \xi^4$, of which there are n, and terms of the form $E[X_i^2 X_j^2] = E[X_i^2]E[X_j^2] = \sigma^4$ for $i \ne j$, of which there are $3n(n-1)$. Hence

(6.2)
$$E[S_n^4] = n\xi^4 + 3n(n-1)\sigma^4 \le Kn^2,$$

where K does not depend on n.

By Markov's inequality (5.31) for $k = 4$, $P[|S_n| \ge n\epsilon] \le Kn^{-2}\epsilon^{-4}$, and so by the first Borel–Cantelli lemma, $P[|n^{-1}S_n| \ge \epsilon \text{ i.o.}] = 0$, as required. ∎

Example 6.1. The classical example is the strong law of large numbers for Bernoulli trials. Here $P[X_n = 1] = p$, $P[X_n = 0] = 1 - p$, $m = p$; S_n represents the number of successes in n trials, and $n^{-1}S_n \to p$ with probability 1. The idea of probability as frequency depends on the long-range stability of the success ratio S_n/n. ∎

Example 6.2. Theorem 1.2 is the case of Example 6.1 in which $(\Omega, \mathscr{F}, P)$ is the unit interval and the $X_n(\omega)$ are the digits $d_n(\omega)$ of the dyadic expansion of ω. Here $p = \frac{1}{2}$. The set (1.21) of normal numbers in the unit interval has by (6.1) Lebesgue measure 1; its complement has measure 0 (and so in the terminology of Section 1 is negligible). ∎

The Weak Law

Since convergence with probability 1 implies convergence in probability (Theorem 5.2(ii)), it follows under the hypotheses of Theorem 6.1 that $n^{-1}S_n \to_P m$. But this is of course an immediate consequence of Chebyshev's inequality (5.32) and the rule (5.28) for adding variances:

$$P\left[\left|n^{-1}S_n - m\right| \geq \epsilon\right] \leq \frac{\mathrm{Var}[S_n]}{n^2\epsilon^2} = \frac{n\,\mathrm{Var}[X_1]}{n^2\epsilon^2} \to 0.$$

This is the *weak law of large numbers.*

Chebyshev's inequality leads to a weak law in other interesting cases as well:

Example 6.3. Let Ω_n consist of the $n!$ permutations of $1, 2, \ldots, n$, all equally probable, and let $X_{nk}(\omega)$ be 1 or 0 according as the kth element in the cyclic representation of $\omega \in \Omega_n$ completes a cycle or not. This is Example 5.6, although there the dependence on n was suppressed in the notation. The $X_{n1}, \ldots, X_{nn}$ are independent, and $S_n = X_{n1} + \cdots + X_{nn}$ is the number of cycles. The mean m_{nk} of X_{nk} is the probability that it equals 1, namely $(n - k + 1)^{-1}$, and its variance is $\sigma_{nk}^2 = m_{nk}(1 - m_{nk})$.

If $L_n = \sum_{k=1}^{n} k^{-1}$, then S_n has mean $\sum_{k=1}^{n} m_{nk} = L_n$ and variance $\sum_{k=1}^{n} m_{nk}(1 - m_{nk}) < L_n$. By Chebyshev's inequality,

$$P\left[\left|\frac{S_n - L_n}{L_n}\right| \geq \epsilon\right] < \frac{L_n}{\epsilon^2 L_n^2} = \frac{1}{\epsilon^2 L_n} \to 0.$$

Of the $n!$ permutations on n letters, a proportion exceeding $1 - \epsilon^{-2}L_n^{-1}$ thus have their cycle number in the range $(1 \pm \epsilon)L_n$. Since $L_n = \log n + O(1)$, most permutations on n letters have about $\log n$ cycles. For a refinement, see Example 27.3.

Since Ω_n changes with n, it is the nature of the case that there cannot be a strong law corresponding to this result. ∎

Bernstein's Theorem

Some theorems that can be stated without reference to probability nonetheless have simple probabilistic proofs, as the last example shows. Bernstein's approach to the Weierstrass approximation theorem is another example.

Let f be a function on $[0, 1]$. The *Bernstein polynomial* of degree n associated with f is

$$(6.3) \qquad B_n(x) = \sum_{k=0}^{n} f\left(\frac{k}{n}\right) \binom{n}{k} x^k (1-x)^{n-k}$$

Theorem 6.2. *If f is continuous, $B_n(x)$ converges to $f(x)$ uniformly on $[0, 1]$.*

According to the Weierstrass approximation theorem, f can be uniformly approximated by polynomials; Bernstein's result goes further and specifies an approximating sequence.

PROOF. Let $M = \sup_x |f(x)|$, and let $\delta(\epsilon) = \sup[|f(x) - f(y)|: |x - y| \le \epsilon]$ be the modulus of continuity of f. It will be shown that

$$(6.4) \qquad \sup_x |f(x) - B_n(x)| \le \delta(\epsilon) + \frac{2M}{n\epsilon^2}.$$

By the uniform continuity of f, $\lim_{\epsilon \to 0} \delta(\epsilon) = 0$, and so this inequality (for $\epsilon = n^{-1/3}$, say) will give the theorem.

Fix $n \ge 1$ and $x \in [0, 1]$ for the moment. Let $X_1, \ldots, X_n$ be independent random variables (on some probability space) such that $P[X_i = 1] = x$ and $P[X_i = 0] = 1 - x$; put $S = X_1 + \cdots + X_n$. Since $P[S = k] = \binom{n}{k} x^k (1-x)^{n-k}$, the formula (5.19) for calculating expected values of functions of random variables gives $E[f(S/n)] = B_n(x)$. By the law of large numbers, there should be high probability that S/n is near x and hence (f being continuous) that $f(S/n)$ is near $f(x)$; $E[f(S/n)]$ should therefore be near $f(x)$. This is the probabilistic idea behind the proof and, indeed, behind the definition (6.3) itself.

Bound $|f(n^{-1}S) - f(x)|$ by $\delta(\epsilon)$ on the set $[|n^{-1}S - x| < \epsilon]$ and by $2M$ on the complementary set, and use (5.22) as in the proof of Theorem 5.4. Since $E[S] = nx$, Chebyshev's inequality gives

$$|B_n(x) - f(x)| \le E\big[|f(n^{-1}S) - f(x)|\big]$$
$$\le \delta(\epsilon) P[|n^{-1}S - x| < \epsilon] + 2MP[|n^{-1}S - x| \ge \epsilon]$$
$$\le \delta(\epsilon) + 2M \operatorname{Var}[S]/n^2\epsilon^2;$$

since $\operatorname{Var}[S] = nx(1 - x) \le n$, (6.4) follows. ∎

A Refinement of the Second Borel–Cantelli Lemma

For a sequence $A_1, A_2, \ldots$ of events, consider the number $N_n = I_{A_1} + \cdots + I_{A_n}$ of occurrences among $A_1, \ldots, A_n$. Since $[A_n \text{ i.o.}] = [\omega: \sup_n N_n(\omega) = \infty]$, $P[A_n \text{ i.o.}]$ can be studied by means of the random variables N_n.

Suppose that the A_n are independent. Put $p_k = P(A_k)$ and $m_n = p_1 + \cdots + p_n$. From $E[I_{A_k}] = p_k$ and $\text{Var}[I_{A_k}] = p_k(1 - p_k) \le p_k$ follow $E[N_n] = m_n$ and $\text{Var}[N_n] = \sum_{k=1}^{n} \text{Var}[I_{A_k}] \le m_n$. If $m_n > x$, then

$$(6.5) \qquad P[N_n \le x] \le P[|N_n - m_n| \ge m_n - x]$$

$$\le \frac{\text{Var}[N_n]}{(m_n - x)^2} \le \frac{m_n}{(m_n - x)^2}.$$

If $\sum p_n = \infty$, so that $m_n \to \infty$, it follows that $\lim_n P[N_n \le x] = 0$ for each x. Since

$$(6.6) \qquad P\left[\sup_k N_k \le x \right] \le P[N_n \le x],$$

$P[\sup_k N_k \le x] = 0$ and hence (take the union over $x = 1, 2, \dots$) $P[\sup_k N_k < \infty] = 0$. Thus $P[A_n \text{ i.o.}] = P[\sup_n N_i = \infty] = 1$ if the A_n are independent and $\sum p_n = \infty$, which proves the second Borel–Cantelli lemma once again.

Independence was used in this argument only to estimate $\text{Var}[N_n]$. Even without independence, $E[N_n] = m_n$ and the first two inequalities in (6.5) hold.

Theorem 6.3. *If $\sum P(A_n)$ diverges and*

$$(6.7) \qquad \liminf_n \frac{\sum_{j, k \le n} P(A_j \cap A_k)}{\left(\sum_{k \le n} P(A_k) \right)^2} \le 1,$$

then $P[A_n \text{ i.o.}] = 1$.

As the proof will show, the ratio in (6.7) is at least 1; if (6.7) holds, the inequality must therefore be an equality.

PROOF. Let θ_n denote the ratio in (6.7). In the notation above,

$$\text{Var}[N_n] = E[N_n^2] - m_n^2 = \sum_{j, k \le n} E[I_{A_j} I_{A_k}] - m_n^2$$

$$= \sum_{j, k \le n} P(A_j \cap A_k) - m_n^2 = (\theta_n - 1)m_n^2$$

(and $\theta_n - 1 \ge 0$). Hence (see (6.5)) $P[N_n \le x] \le (\theta_n - 1)m_n^2/(m_n - x)^2$ for $x < m_n$. Since $m_n^2/(m_n - x)^2 \to 1$, (6.7) implies that $\liminf_n P[N_n \le x] = 0$. It still follows by (6.6) that $P[\sup_k N_k \le x] = 0$, and the rest of the argument is as before. ∎

Example 6.4.   If, as in the second Borel–Cantelli lemma, the A_n are independent (or even if they are merely independent in pairs), the ratio in (6.7) is $1 + \sum_{k \leq n}(p_k - p_k^2)/m_n^2$, so that $\sum P(A_n) = \infty$ implies (6.7).   ·   ∎

Example 6.5.   Return once again to the run lengths $l_n(\omega)$ of Section 4. It was shown in Example 4.21 that $\{r_n\}$ is an outer boundary $(P[l_n \geq r_n \text{ i.o.}] = 0)$ if $\sum 2^{-r_n} < \infty$. It was also shown that $\{r_n\}$ is an inner boundary $(P[l_n \geq r_n \text{ i.o.}] = 1)$ if r_n is nondecreasing and $\sum 2^{-r_n}r_n^{-1} = \infty$, but Theorem 6.3 can be used to prove this under the sole assumption that $\sum 2^{-r_n} = \infty$.

As usual, the r_n can be taken to be positive integers. Let $A_n = [l_n \geq r_n] = [d_n = \cdots = d_{n+r_n-1} = 0]$. If $j + r_j \leq k$, then A_j and A_k are independent. If $j < k < j + r_j$, then $P(A_j|A_k) \leq P[d_j = \cdots = d_{k-1} = 0|A_k] = P[d_j = \cdots = d_{k-1} = 0] = 1/2^{k-j}$, and so $P(A_j \cap A_k) \leq P(A_k)/2^{k-j}$. Therefore,

$$\sum_{j,\,k \leq n} P(A_j \cap A_k)$$

$$\leq \sum_{k \leq n} P(A_k) + 2 \sum_{\substack{j < k \leq n \\ j + r_j \leq k}} P(A_j)P(A_k) + 2 \sum_{\substack{j < k \leq n \\ k < j + r_j}} 2^{-(k-j)}P(A_k)$$

$$\leq \sum_{k \leq n} P(A_k) + \left(\sum_{k \leq n} P(A_k) \right)^2 + 2 \sum_{k \leq n} P(A_k).$$

If $\sum P(A_n) = \sum 2^{-r_n}$ diverges, then (6.7) follows.

Thus $\{r_n\}$ *is an outer or an inner boundary according as* $\sum 2^{-r_n}$ *converges or diverges*, which completely settles the issue. In particular, $r_n = \log_2 n + \theta \log_2 \log_2 n$ gives an outer boundary for $\theta > 1$ and an inner boundary for $\theta \leq 1$.   ∎

Example 6.6.   It is now possible to complete the analysis in Examples 4.12 and 4.16 of the relative error $e_n(\omega)$ in the approximation of ω by $\sum_{k=1}^{n-1} d_k(\omega)2^{-k}$. If $l_n(\omega) \geq -\log_2 x_n$ $(0 < x_n < 1)$, then $e_n(\omega) \leq x_n$ by (4.22). By the preceding example for the case $r_n = -\log_2 x_n$, $\sum x_n = \infty$ implies that $P[\omega: e_n(\omega) \leq x_n \text{ i.o.}] = 1$. By this and Example 4.12, $[\omega: e_n(\omega) \leq x_n \text{ i.o.}]$ *has Lebesgue measure* 0 *or* 1 *according as* $\sum x_n$ *converges or diverges.*   ∎

PROBLEMS

6.1. Show that $Z_n \to Z$ with probability 1 if and only if for every positive ϵ there exists an n such that $P[|Z_k - Z| < \epsilon, n \leq k \leq m] > 1 - \epsilon$ for all m exceeding n. This describes convergence with probability 1 in "finite" terms.

6.2. Show in Example 6.3 that $P[|S_n - L_n| \geq L_n^{1/2+\epsilon}] \to 0$.

6.3. As in Examples 5.6 and 6.3, let ω be a random permutation of $1, 2, \ldots, n$ Each k, $1 \leq k \leq n$, occupies some position in the bottom row of the permutation ω;

let $X_{nk}(\omega)$ be the number of smaller elements (between 1 and $k-1$) lying to the right of k in the bottom row. The sum $S_n = X_{n1} + \cdots + X_{nn}$ is the total number of *inversions*—the number of pairs appearing in the bottom row in reverse order of size. For the permutation in Example 5.6 the values of $X_{71}, \ldots, X_{77}$ are 0, 0, 0, 2, 4, 2, 4, and $S_7 = 12$. Show that $X_{n1}, \ldots, X_{nn}$ are independent and $P[X_{nk} = i] = k^{-1}$ for $0 \le i < k$. Calculate $E[S_n]$ and $Var[S_n]$. Show that S_n is likely to be near $n^2/4$.

6.4. For a function f on $[0,1]$ write $\|f\| = \sup_x |f(x)|$. Show that, if f has a continuous derivative f', then $\|f - B_n\| \le \epsilon \|f'\| + 2\|f\|/n\epsilon^2$. Conclude that $\|f - B_n\| = O(n^{-1/3})$.

6.5. Prove *Poisson's theorem*: If $A_1, A_2, \ldots$ are independent events, $\bar{p}_n = n^{-1}\sum_{i=1}^{n} P(A_i)$, and $N = \sum_{i=1}^{n} I_{A_i}$, then $n^{-1}N_n - \bar{p}_n \to_p 0$.

In the following problems $S_n = X_1 + \cdots + X_n$

6.6. Prove *Cantelli's theorem*. If $X_1, X_2, \ldots$ are independent, $E[X_n] = 0$, and $E[X_n^4]$ is bounded, then $n^{-1}S_n \to 0$ with probability 1. The X_n need not be identically distributed

6.7. (a) Let $x_1, x_2, \ldots$ be a sequence of real numbers, and put $s_n = x_1 + \cdots + x_n$. Suppose that $n^{-2}s_{n^2} \to 0$ and that the x_n are bounded, and show that $n^{-1}s_n \to 0$.
(b) Suppose that $n^{-2}S_{n^2} \to 0$ with probability 1 and that the X_n are uniformly bounded ($\sup_{n,\omega} |X_n(\omega)| < \infty$). Show that $n^{-1}S_n \to 0$ with probability 1 Here the X_n need not be identically distributed or even independent.

6.8. ↑ Suppose that $X_1, X_2, \ldots$ are independent and uniformly bounded and $E[X_n] = 0$. Using only the preceding result, the first Borel–Cantelli lemma, and Chebyshev's inequality, prove that $n^{-1}S_n \to 0$ with probability 1.

6.9. ↑ Use the ideas of Problem 6.8 to give a new proof of Borel's normal number theorem, Theorem 1.2. The point is to return to first principles and use only negligibility and the other ideas of Section 1, not the apparatus of Sections 2 through 6; in particular, $P(A)$ is to be taken as defined only if A is a finite, disjoint union of intervals.

6.10. 5.11 6.7↑ Suppose that (in the notation of (5.41)) $\beta_n - \alpha_n^2 = O(1/n)$. Show that $n^{-1}N_n - \alpha_n \to 0$ with probability 1. What condition on $\beta_n - \alpha_n^2$ will imply a weak law? Note that independence is not assumed here.

6.11. Suppose that $X_1, X_2, \ldots$ are *m-dependent* in the sense that random variables more than m apart in the sequence are independent. More precisely, let $\mathscr{A}_j^k = \sigma(X_j, \ldots, X_k)$, and assume that $\mathscr{A}_{j_1}^{k_1}, \ldots, \mathscr{A}_{j_l}^{k_l}$ are independent if $k_{i-1} + m < j_i$ for $i = 2, \ldots, l$. (Independent random variables are 0-dependent.) Suppose that the X_n have this property and are uniformly bounded and that $E[X_n] = 0$. Show that $n^{-1}S_n \to 0$. *Hint*: Consider the subsequences $X_i, X_{i+m+1}, X_{i+2(m+1)},$ for $1 \le i \le m+1$.

6.12. ↑ Suppose that the X_n are independent and assume the values $x_1, \ldots, x_l$ with probabilities $p(x_1), \ldots, p(x_l)$. For $u_1, \ldots, u_k$ a k-tuple of the x_i's, let

$N_n(u_1,\ldots,u_k)$ be the frequency of the k-tuple in the first $n + k - 1$ trials, that is, the number of t such that $1 \le t \le n$ and $X_t = u_1, \ldots, X_{t+k-1} = u_k$. Show that with probability 1, all asymptotic relative frequencies are what they should be—that is, with probability 1, $n^{-1}N_n(u_1,\ldots,u_k) \to p(u_1)\cdots p(u_k)$ for every k and every k-tuple $u_1,\ldots,u_k$.

6.13. ↑   A number ω in the unit interval is *completely normal* if, for every base b and every k and every k-tuple of base-b digits, the k-tuple appears in the base-b expansion of ω with asymptotic relative frequency b^{-k}. Show that the set of completely normal numbers has Lebesgue measure 1.

6.14. *Shannon's theorem.* Suppose that $X_1, X_2,\ldots$ are independent, identically distributed random variables taking on the values $1,\ldots,r$ with positive probabilities $p_1,\ldots,p_r$. If $p_n(i_1,\ldots,i_n) = p_{i_1}\cdots p_{i_n}$ and $p_n(\omega) = p_n(X_1(\omega),\ldots,X_n(\omega))$, then $p_n(\omega)$ is the probability that a new sequence of n trials would produce the particular sequence $X_1(\omega),\ldots,X_n(\omega)$ of outcomes that happens actually to have been observed. Show that

$$-\frac{1}{n}\log p_n(\omega) \to h = -\sum_{i=1}^{r} p_i \log p_i$$

with probability 1.

In information theory $1,\ldots,r$ are interpreted as the *letters* of an *alphabet*, $X_1, X_2,\ldots$ are the successive letters produced by an information *source*, and h is the *entropy* of the source. Prove the *asymptotic equipartition property*: For large n there is probability exceeding $1 - \epsilon$ that the probability $p_n(\omega)$ of the observed n-long sequence, or *message*, is in the range $e^{-n(h \pm \epsilon)}$.

6.15. In the terminology of Example 6.5, show that $\log_2 n + \log_2 \log_2 n + \theta \log_2 \log_2 \log_2 n$ is an outer or inner boundary as $\theta > 1$ or $\theta \le 1$. Generalize. (Compare Problem 4.12.)

6.16. 5.20↑   Let $g(m) = \sum_p \delta_p(m)$ be the number of distinct prime divisors of m. For $a_n = E_n[g]$ (see (5.46)) show that $a_n \to \infty$. Show that

$$(6.8) \qquad E_n\left[\left(\delta_p - \frac{1}{n}\left\lfloor\frac{n}{p}\right\rfloor\right)\left(\delta_q - \frac{1}{n}\left\lfloor\frac{n}{q}\right\rfloor\right)\right] \le \frac{1}{np} + \frac{1}{nq}$$

for $p \ne q$ and hence that the variance of g under P_n satisfies

$$(6.9) \qquad \mathrm{Var}_n[g] \le 3 \sum_{p \le n} \frac{1}{p}.$$

Prove the *Hardy–Ramanujan theorem*:

$$(6.10) \qquad \lim_n P_n\left[m: \left|\frac{g(m)}{a_n} - 1\right| \ge \epsilon\right] = 0.$$

Since $a_n \sim \log\log n$ (see Problem 18.17), most integers under n have something like $\log\log n$ distinct prime divisors. Since $\log\log 10^7$ is a little less than 3, the typical integer under 10^7 has about three prime factors—remarkably few.

6.17. Suppose that $X_1, X_2, \ldots$ are independent and $P[X_n = 0] = p$. Let L_n be the length of the run of 0's starting at the nth place: $L_n = k$ if $X_n = \cdots = X_{n+k-1} = 0 \neq X_{n+k}$. Show that $P[L_n \geq r_n \text{ i.o.}]$ is 0 or 1 according as $\sum_n p^{r_n}$ converges or diverges. Example 6.5 covers the case $p = \frac{1}{2}$.

SECTION 7. GAMBLING SYSTEMS

Let $X_1, X_2, \ldots$ be an independent sequence of random variables (on some $(\Omega, \mathscr{F}, P)$) taking on the two values $+1$ and -1 with probabilities $P[X_n = +1] = p$ and $P[X_n = -1] = q = 1 - p$. Throughout the section, X_n will be viewed as the gambler's gain on the nth of a series of plays at unit stakes. The game is favorable to the gambler if $p > \frac{1}{2}$, fair if $p = \frac{1}{2}$, and unfavorable if $p < \frac{1}{2}$. The case $p \leq \frac{1}{2}$ will be called the *subfair case*.

After the classical gambler's ruin problem has been solved, it will be shown that every gambling system is in certain respects without effect and that some gambling systems are in other respects optimal. Gambling problems of the sort considered here have inspired many ideas in the mathematical theory of probability, ideas that carry far beyond their origin.

Red-and-black will provide numerical examples. Of the 38 spaces on a roulette wheel, 18 are red, 18 are black, and 2 are green. In betting either on red or on black the chance of winning is $\frac{18}{38}$.

Gambler's Ruin

Suppose that the gambler enters the casino with capital a and adopts the strategy of continuing to bet at unit stakes until his fortune increases to c or his funds are exhausted. What is the probability of ruin, the probability that he will lose his capital, a? What is the probability he will achieve his goal, c? Here a and c are integers.

Let

$$(7.1) \qquad\qquad S_n = X_1 + \cdots + X_n, \qquad S_0 = 0.$$

The gambler's fortune after n plays is $a + S_n$. The event

$$(7.2) \qquad A_{a,n} = [a + S_n = c] \cap \bigcap_{k=1}^{n-1} [0 < a + S_k < c]$$

represents success for the gambler at time n, and

$$(7.3) \qquad B_{a,n} = [a + S_n = 0] \cap \bigcap_{k=1}^{n-1} [0 < a + S_k < c]$$

represents ruin at time n. If $s_c(a)$ denotes the probability of ultimate success, then

$$(7.4) \qquad s_c(a) = P\left(\bigcup_{n=1}^{\infty} A_{a,n} \right) = \sum_{n=1}^{\infty} P(A_{a,n})$$

for $0 < a < c$.

Fix c and let a vary. For $n \geq 1$ and $0 < a < c$, define $A_{a,n}$ by (7.2), and adopt the conventions $A_{a,0} = \varnothing$ for $0 \leq a < c$ and $A_{c,0} = \Omega$ (success is impossible at time 0 if $a < c$ and certain if $a = c$), as well as $A_{0,n} = A_{c,n} = \varnothing$ for $n \geq 1$ (play never starts if a is 0 or c). By these conventions, $s_c(0) = 0$ and $s_c(c) = 1$.

Because of independence and the fact that the sequence $X_2, X_3, \ldots$ is a probabilistic replica of $X_1, X_2, \ldots$, it seems clear that the chance of success for a gambler with initial fortune a must be the chance of winning the first wager times the chance of success for an initial fortune $a + 1$, plus the chance of losing the first wager times the chance of success for an initial fortune $a - 1$. It thus seems intuitively clear that

$$(7.5) \qquad s_c(a) = p s_c(a + 1) + q s_c(a - 1), \qquad 0 < a < c.$$

For a rigorous argument, define $A'_{a,n}$ just as $A_{a,n}$ but with $S'_n = X_2 + \cdots + X_{n+1}$ in place of S_n in (7.2). Now $P[X_i = x_i,\ i \leq n] = P[X_{i+1} = x_i,\ i \leq n]$ for each sequence $x_1, \ldots, x_n$ of $+1$'s and -1's, and therefore $P[(X_1, \ldots, X_n) \in H] = P[(X_2, \ldots, X_{n+1}) \in H]$ for $H \subset R^n$. Take H to be the set of $x = (x_1, \ldots, x_n)$ in R^n satisfying $x_i = \pm 1$, $a + x_1 + \cdots + x_n = c$, and $0 < a + x_1 + \cdots + x_k < c$ for $k < n$. It follows then that

$$(7.6) \qquad P(A_{a,n}) = P(A'_{a,n}).$$

Moreover, $A_{a,n} = ([X_1 = +1] \cap A'_{a+1,n-1}) \cup ([X_1 = -1] \cap A'_{a-1,n-1})$ for $n \geq 1$ and $0 < a < c$. By independence and (7.6), $P(A_{a,n}) = pP(A_{a+1,n-1}) + qP(A_{a-1,n-1})$; adding over n now gives (7.5). Note that this argument involves the entire infinite sequence $X_1, X_2, \ldots$.

It remains to solve the difference equation (7.5) with the side conditions $s_c(0) = 0$, $s_c(c) = 1$. Let $\rho = q/p$ be the odds against the gambler. Then [A19] there exist constants A and B such that, for $0 \leq a \leq c$, $s_c(a) = A + B\rho^a$ if $p \neq q$ and $s_c(a) = A + Ba$ if $p = q$. The requirements $s_c(0) = 0$ and $s_c(c) = 1$ determine A and B, which gives the solution:

The probability that the gambler can before ruin attain his goal of c from an initial capital of a is

$$(7.7) \qquad s_c(a) = \begin{cases} \dfrac{\rho^a - 1}{\rho^c - 1}, & 0 \leq a \leq c, \quad \text{if } \rho = \dfrac{q}{p} \neq 1, \\[2ex] \dfrac{a}{c}, & 0 \leq a \leq c, \quad \text{if } \rho = \dfrac{q}{p} = 1. \end{cases}$$

Example 7.1. The gambler's initial capital is \$900 and his goal is \$1000. If $p = \frac{1}{2}$, his chance of success is very good: $s_{1000}(900) = .9$. At red-and-black, $p = \frac{18}{38}$ and hence $\rho = \frac{20}{18}$; in this case his chance of success as computed by (7.7) is only about .00003.   ∎

Example 7.2. It is the gambler's desperate intention to convert his \$100 into \$20,000. For a game in which $p = \frac{1}{2}$ (no casino has one), his chance of success is $100/20,000 = .005$; at red-and-black it is minute—about 3×10^{-911}

∎

In the analysis leading to (7.7), replace (7.2) by (7.3). It follows that (7.7) with p and q interchanged (ρ goes to ρ^{-1}) and a and $c - a$ interchanged gives the probability $r_c(a)$ of ruin for the gambler: $r_c(a) = (\rho^{-(c-a)} - 1)/(\rho^{-c} - 1)$ if $\rho \neq 1$ and $r_c(a) = (c - a)/c$ if $\rho = 1$. Hence $s_c(a) + r_c(a) = 1$ holds in all cases: *The probability is 0 that play continues forever*.

For positive integers a and b, let

$$H_{a,b} = \bigcup_{n=1}^{\infty} \left\{ [S_n = b] \cap \bigcap_{k=1}^{n-1} [-a < S_k < b] \right\}$$

be the event that S_n reaches $+b$ before reaching $-a$. Its probability is simply (7.7) with $c = a + b$: $P(H_{a,b}) = s_{a+b}(a)$. Now let

$$H_b = \bigcup_{a=1}^{\infty} H_{a,b} = \bigcup_{n=1}^{\infty} [S_n = b] = \left[\sup_n S_n \geq b \right]$$

be the event that S_n ever reaches $+b$. Since $H_{a,b} \uparrow H_b$ as $a \to \infty$, it follows that $P(H_b) = \lim_a s_{a+b}(a)$; this is 1 if $\rho = 1$ or $\rho < 1$, and it is $1/\rho^b$ if $\rho > 1$. Thus

$$(7.8) \qquad\qquad P\left[\sup_n S_n \geq b \right] = \begin{cases} 1 & \text{if } p \geq q, \\ (p/q)^b & \text{if } p < q. \end{cases}$$

This is the probability that a gambler with unlimited capital can ultimately gain b units.

Example 7.3. The gambler in Example 7.1 has capital 900 and the goal of winning $b = 100$; in Example 7.2 he has capital 100 and b is 19,900. Suppose, instead, that his capital is infinite. If $p = \frac{1}{2}$, the chance of achieving his goal increases from .9 to 1 in the first example and from .005 to 1 in the second. At red-and-black, however, the two probabilities $.9^{100}$ and $.9^{19900}$ remain essentially what they were before (.00003 and 3×10^{-911}).   ∎

Selection Systems

Players often try to improve their luck by betting only when in the preceding trials the wins and losses form an auspicious pattern. Perhaps the gambler bets on the nth trial only when among $X_1, \ldots, X_{n-1}$ there are many more $+1$'s than -1's, the idea being to ride winning streaks (he is "in the vein"). Or he may bet only when there are many more -1's than $+1$'s, the idea being it is then surely time a $+1$ came along (the "maturity of the chances"). There is a mathematical theorem that, translated into gaming language, says all such systems are futile.

It might be argued that it *is* sensible to bet if among $X_1, \ldots, X_{n-1}$ there is an excess of $+1$'s, on the ground that it is evidence of a high value of p. But it is assumed throughout that statistical inference is not at issue: p is fixed—at $\frac{18}{38}$, for example, in the case of red-and-black—and is known to the gambler, or should be.

The gambler's strategy is described by random variables $B_1, B_2, \ldots$ taking the two values 0 and 1: If $B_n = 1$, the gambler places a bet on the nth trial; if $B_n = 0$, he skips that trial. If B_n were $(X_n + 1)/2$, so that $B_n = 1$ for $X_n = +1$ and $B_n = 0$ for $X_n = -1$, the gambler would win every time he bet, but of course such a system requires he be prescient—he must know the outcome X_n in advance. For this reason the value of B_n is assumed to depend only on the values of $X_1, \ldots, X_{n-1}$: there exists some function $b_n \colon R^{n-1} \to R^1$ such that

$$(7.9) \qquad B_n = b_n(X_1, \ldots, X_{n-1}).$$

(Here B_1 is constant.) Thus the mathematics avoids, as it must, the question of whether prescience is actually possible.

Define

$$(7.10) \qquad \begin{cases} \mathscr{F}_n = \sigma(X_1, \ldots, X_n), & n = 1, 2, \ldots; \\ \mathscr{F}_0 = \{\varnothing, \Omega\}. \end{cases}$$

The σ-field $\mathscr{F}_{n-1}$ generated by $X_1, \ldots, X_{n-1}$ corresponds to a knowledge of the outcomes of the first $n - 1$ trials. The requirement (7.9) ensures that B_n is measurable $\mathscr{F}_{n-1}$ (Theorem 5.1) and so depends only on the information actually available to the gambler just before the nth trial.

For $n = 1, 2, \ldots$, let N_n be the time at which the gambler places his nth bet. This nth bet is placed at time k or earlier if and only if the number $\sum_{i=1}^{k} B_i$ of bets placed up to and including time k is n or more; in fact, N_n is the smallest k for which $\sum_{i=1}^{k} B_i = n$. Thus the event $[N_n \leq k]$ coincides with $[\sum_{i=1}^{k} B_i \geq n]$; by (7.9) this latter event lies in $\sigma(B_1, \ldots, B_k) \subset \sigma(X_1, \ldots, X_{k-1}) = \mathscr{F}_{k-1}$. Therefore,

$$(7.11) \qquad [N_n = k] = [N_n \leq k] - [N_n \leq k - 1] \in \mathscr{F}_{k-1}.$$

(Even though $[N_n = k]$ lies in $\mathcal{F}_{k-1}$ and hence in $\mathcal{F}$, N_n is, as a function on Ω, generally not a simple random variable, because it has infinite range. This makes no difference, because expected values of the N_n will play no role; (7.11) is the essential property.)

To ensure that play continues forever (stopping rules will be considered later) and that the N_n have finite values with probability 1, make the further assumption that

$$(7.12) \qquad\qquad P[B_n = 1 \text{ i.o.}] = 1.$$

A sequence $\{B_n\}$ of random variables assuming the values 0 and 1, having the form (7.9), and satisfying (7.12) is a *selection system*.

Let Y_n be the gambler's gain on the nth of the trials at which he does bet: $Y_n = X_{N_n}$. It is only on the set $[B_n = 1 \text{ i.o}]$ that all the N_n and hence all the Y_n are well defined. To complete the definition, set $Y_n = -1$, say, on $[B_n = 1$ i.o.$]^c$; since this set has probability 0 by (7.12), it really makes no difference how Y_n is defined on it.

Now Y_n is a complicated function on Ω because $Y_n(\omega) = X_{N_n(\omega)}(\omega)$. Nonetheless,

$$[\omega: Y_n(\omega) = 1] = \bigcup_{k=1}^{\infty} \left([\omega: N_n(\omega) = k] \cap [\omega: X_k(\omega) = 1] \right)$$

lies in $\mathcal{F}$, and so does its complement $[\omega: Y_n(\omega) = -1]$. Hence Y_n is a simple random variable.

Example 7.4. An example will fix these ideas. Suppose that the rule is always to bet on the first trail, to bet on the second trial if and only if $X_1 = +1$, to bet on the third trial if and only if $X_1 = X_2$, and to bet on all subsequent trails. Here $B_1 = 1$, $[B_2 = 1] = [X_1 = +1]$, $[B_3 = 1] = [X_1 = X_2]$, and $B_4 = B_5 = \cdots = 1$. The table shows the ways the gambling can start out. A dot represents a value undetermined by X_1, X_2, X_3. Ignore the rightmost column for the moment.

X_1	X_2	X_3	B_1	B_2	B_3	N_1	N_2	N_3	N_4	Y_1	Y_2	Y_3	τ
-1	-1	-1	1	0	1	1	3	4	5	-1	-1	$\cdot$	1
-1	-1	$+1$	1	0	1	1	3	4	5	-1	$+1$	$\cdot$	1
-1	$+1$	-1	1	0	0	1	4	5	6	-1	$\cdot$	$\cdot$	1
-1	$+1$	$+1$	1	0	0	1	4	5	6	-1	$\cdot$	$\cdot$	1
$+1$	-1	-1	1	1	0	1	2	4	5	$+1$	-1	$\cdot$	2
$+1$	-1	$+1$	1	1	0	1	2	4	5	$+1$	-1	$\cdot$	2
$+1$	$+1$	-1	1	1	1	1	2	3	4	$+1$	$+1$	-1	3
$+1$	$+1$	$+1$	1	1	1	1	2	3	4	$+1$	$+1$	$+1$	$\cdot$

In the evolution represented by the first line of the table, the second bet is placed on the third trial ($N_2 = 3$), which results in a loss because $Y_2 = X_{N_2} = X_3 = -1$. Since $X_3 = -1$, the gambler was "wrong" to bet. But remember that before the third trial he does not know $X_3(\omega)$ (much less ω itself); he knows only $X_1(\omega)$ and $X_2(\omega)$. See the discussion in Example 5.5. ∎

Selection systems achieve nothing because $\{Y_n\}$ has the same structure as $\{X_n\}$:

Theorem 7.1. *For every selection system, $\{Y_n\}$ is independent and $P[Y_n = +1] = p$, $P[Y_n = -1] = q$.*

PROOF. Since random variables with indices that are themselves random variables are conceptually confusing at first, the ω's here will not be suppressed as they have been in previous proofs.

Relabel p and q as $p(+1)$ and $p(-1)$, so that $P[\omega: X_k(\omega) = x] = p(x)$ for $x = \pm 1$. If $A \in \mathscr{F}_{k-1}$, then A and $[\omega: X_k(\omega) = x]$ are independent, and so $P(A \cap [\omega: X_k(\omega) = x]) = P(A)p(x)$. Therefore, by (7.11),

$$P\big[\omega: Y_n(\omega) = x\big] = P\big[\omega: X_{N_n(\omega)}(\omega) = x\big]$$

$$= \sum_{k=1}^{\infty} P\big[\omega: N_n(\omega) = k, X_k(\omega) = x\big]$$

$$= \sum_{k=1}^{\infty} P\big[\omega: N_n(\omega) = k\big] p(x)$$

$$= p(x).$$

More generally, for any sequence $x_1, \ldots, x_n$ of ± 1's,

$$P\big[\omega: Y_i(\omega) = x_i, i \le n\big] = P\big[\omega: X_{N_i(\omega)}(\omega) = x_i, i \le n\big]$$

$$= \sum_{k_1 < \cdots < k_n} P\big[\omega: N_i(\omega) = k_i, X_{k_i}(\omega) = x_i, i \le n\big],$$

where the sum extends over n-tuples of positive integers satisfying $k_1 < \cdots < k_n$. The event $[\omega: N_i(\omega) = k_i, i \le n] \cap [\omega: X_{k_i}(\omega) = x_i, i < n]$ lies in $\mathscr{F}_{k_n - 1}$ (note that there is no condition on $X_{k_n}(\omega)$), and therefore

$$P\big[\omega: Y_i(\omega) = x_i, i \le n\big]$$

$$= \sum_{k_1 < \quad < k_n} P\big([\omega: N_i(\omega) = k_i, i \le n]$$

$$\cap [\omega: X_{k_i}(\omega) = x_i, i < n]\big) p(x_n).$$

Summing k_n over $k_{n-1} + 1$, $k_{n-1} + 2, \ldots$ brings this last sum to

$$\sum_{k_1 < \cdots < k_{n-1}} P\big[\omega\colon N_i(\omega) = k_i,\, X_{k_i}(\omega) = x_i,\, i < n\big] p(x_n)$$

$$= P\big[\omega\colon X_{N_i(\omega)}(\omega) = x_i,\, i < n\big] p(x_n)$$

$$= P\big[\omega\colon Y_i(\omega) = x_i,\, i < n\big] p(x_n).$$

It follows by induction that

$$P\big[\omega\colon Y_i(\omega) = x_i,\, i \le n\big] = \prod_{i \le n} p(x_i) = \prod_{i \le n} P\big[\omega\colon Y_i(\omega) = x_i\big],$$

and so the Y_i are independent (see (5.9)).   ∎

Gambling Policies

There are schemes that go beyond selection systems and tell the gambler not only whether to bet but how much. Gamblers frequently contrive or adopt such schemes in the confident expectation that they can, by pure force of arithmetic, counter the most adverse workings of chance. If the wager specified for the nth trial is in the amount W_n and the gambler cannot see into the future, then W_n must depend only on $X_1, \ldots, X_{n-1}$. Assume therefore that W_n is a nonnegative function of these random variables: there is an $f_n\colon R^{n-1} \to R^1$ such that

$$(7.13) \qquad\qquad W_n = f_n(X_1, \ldots, X_{n-1}) \ge 0.$$

Apart from nonnegativity there are at the outset no constraints on the f_n, although in an actual casino their values must be integral multiples of a basic unit. Such a sequence $\{W_n\}$ is a *betting system*. Since $W_n = 0$ corresponds to a decision not to bet at all, betting systems in effect include selection systems. In the double-or-nothing system, $W_n = 2^{n-1}$ if $X_1 = \cdots = X_{n-1} = -1$ ($W_1 = 1$) and $W_n = 0$ otherwise.

The amount the gambler wins on the nth play is $W_n X_n$. If his fortune at time n is F_n, then

$$(7.14) \qquad\qquad F_n = F_{n-1} + W_n X_n.$$

This also holds for $n = 1$ if F_0 is taken as his initial (nonrandom) fortune. It is convenient to let W_n depend on F_0 as well as the past history of play and hence to generalize (7.13) to

$$(7.15) \qquad\qquad W_n = g_n(F_0, X_1, \ldots, X_{n-1}) \ge 0$$

for a function g_n: $R^n \to R^1$. In expanded notation, $W_n(\omega) = g_n(F_0, X_1(\omega),$
$\ldots, X_{n-1}(\omega))$. The symbol W_n does not show the dependence on ω or on F_0,
either. For each *fixed* initial fortune F_0, W_n is a simple random variable; by
(7.15) it is measurable $\mathscr{F}_{n-1}$. Similarly, F_n is a function of F_0 as well as of
$X_1(\omega), \ldots, X_n(\omega)$: $F_n = F_n(F_0, \omega)$.

If $F_0 = 0$ and $g_n \equiv 1$, the F_n reduce to the partial sums (7.1).

Since $\mathscr{F}_{n-1}$ and $\sigma(X_n)$ are independent, and since W_n is measurable
$\mathscr{F}_{n-1}$ (for each fixed F_0), W_n and X_n are independent. Therefore, $E[W_n X_n]$
$= E[W_n] \cdot E[X_n]$. Now $E[X_n] = p - q \le 0$ in the subfair case ($p \le \frac{1}{2}$), with
equality in the fair case ($p = \frac{1}{2}$). Since $E[W_n] \ge 0$, (7.14) implies that $E[F_n] \le$
$E[F_{n-1}]$. Therefore,

$$(7.16) \qquad F_0 \ge E[F_1] \ge \quad \cdot \cdot \ge E[F_n] \ge \cdots$$

in the subfair case, and

$$(7.17) \qquad F_0 = E[F_1] = \cdots = E[F_n] = \cdots$$

in the fair case. (If $p < q$ and $P[W_n > 0] > 0$, there is strict inequality in
(7.16).) Thus no betting system can convert a subfair game into a profitable
enterprise.

Suppose that in addition to a betting system, the gambler adopts some
policy for quitting. Perhaps he stops when his fortune reaches a set target, or
his funds are exhausted, or the auguries are in some way dissuasive. The
decision to stop must depend only on the initial fortune and the history of
play up to the present.

Let $\tau(F_0, \omega)$ be a nonnegative integer for each ω in Ω and each $F_0 \ge 0$. If
$\tau = n$, the gambler plays on the nth trial (betting W_n) and then stops; if $\tau = 0$,
he does not begin gambling in the first place. The event $[\omega: \tau(F_0, \omega) = n]$
represents the decision to stop just after the nth trial, and so, whatever value
F_0 may have, it must depend only on $X_1, \ldots, X_n$. Therefore, assume that

$$(7.18) \qquad [\omega: \tau(F_0, \omega) = n] \in \mathscr{F}_n, \qquad n = 0, 1, 2, \ldots$$

A τ satisfying this requirement is a *stopping time*. (In general it has infinite
range and hence is not a simple random variable; as expected values of τ
play no role here, this does not matter.) It is technically necessary to let
$\tau(F_0, \omega)$ be undefined or infinite on an ω-set of probability 0. This has no
effect on the requirement (7.18), which must hold for each finite n. But it *is*
assumed that τ *is finite with probability* 1: play is certain to terminate.

A betting system together with a stopping time is a *gambling policy*. Let π
denote such a policy.

Example 7.5. Suppose that the betting system is given by $W_n = B_n$, with
B_n as in Example 7.4. Suppose that the stopping rule is to quit after the first

loss of a wager. Then $[\tau = n] = \bigcup_{k=1}^{n}[N_k = n, Y_1 = \cdots = Y_{k-1} = +1, Y_k = -1]$. For $j \le k \le n$, $[N_k = n, Y_j = x] = \bigcup_{m=1}^{n}[N_k = n, N_j = m, X_m = x]$ lies in $\mathscr{F}_n$ by (7.11); hence τ is a stopping time. The values of τ are shown in the rightmost column of the table.   ∎

The sequence of fortunes is governed by (7.14) until play terminates, and then the fortune remains for all future time fixed at F_τ (with value $F_{\tau(F_0, \omega)}(\omega)$). Therefore, the gambler's fortune at time n is

(7.19)
$$F_n^* = \begin{cases} F_n & \text{if } \tau \ge n, \\ F_\tau & \text{if } \tau \le n. \end{cases}$$

Note that the case $\tau = n$ is covered by both clauses here If $n - 1 < n \le \tau$, then $F_n^* = F_n = F_{n-1} + W_n X_n = F_{n-1}^* + W_n X_n$; if $\tau \le n - 1 < n$, then $F_n^* = F_\tau = F_{n-1}^*$. Therefore, if $W_n^* = I_{[\tau \ge n]} W_n$, then

(7.20)
$$F_n^* = F_{n-1}^* + I_{[\tau \ge n]} W_n X_n = F_{n-1}^* + W_n^* X_n.$$

But this is the equation for a new betting system in which the wager placed at time n is W_n^*. If $\tau \ge n$ (play has not already terminated), W_n^* is the old amount W_n; if $\tau < n$ (play has terminated), W_n^* is 0. Now by (7.18), $[\tau \ge n] = [\tau < n]^c$ lies in $\mathscr{F}_{n-1}$. Thus $I_{[\tau \ge n]}$ is measurable $\mathscr{F}_{n-1}$, so that W_n^* as well as W_n is measurable $\mathscr{F}_{n-1}$, and $\{W_n^*\}$ represents a legitimate betting system. Therefore, (7.16) and (7.17) apply to the new system:

(7.21)
$$F_0 = F_0^* \ge E[F_1^*] \ge \cdots \ge E[F_n^*] \ge \cdots$$

if $p \le \frac{1}{2}$, and

(7.22)
$$F_0 = F_0^* = E[F_1^*] = \cdots = E[F_n^*] = \ldots$$

if $p = \frac{1}{2}$.

The gambler's ultimate fortune is F_τ. Now $\lim_n F_n^* = F_\tau$ with probability 1, since in fact $F_n^* = F_\tau$ for $n \ge \tau$. If

(7.23)
$$\lim_n E[F_n^*] = E[F_\tau],$$

then (7.21) and (7.22), respectively, imply that $E[F_\tau] \le F_0$ and $E[F_\tau] = F_0$. According to Theorem 5.4, (7.23) does hold if the F_n^* are uniformly bounded.

Call the policy *bounded by* M (M nonrandom) if

(7.24)
$$0 \le F_n^* \le M, \qquad n = 0, 1, 2, \ldots .$$

If F_n^* is not bounded above, the gambler's adversary must have infinite capital. A negative F_n^* represents a debt, and if F_n^* is not bounded below,

the gambler must have a patron of infinite wealth and generosity from whom to borrow and so must in effect have infinite capital. In case F_n^* is bounded below, 0 is the convenient lower bound—the gambler is assumed to have in hand all the capital to which he has access. In any real case, (7.24) holds and (7.23) follows. (There is a technical point that arises because the general theory of integration has been postponed: F_τ must be assumed to have finite range so that it will be a simple random variable and hence have an expected value in the sense of Section 5.[†]) The argument has led to this result:

Theorem 7.2. *For every policy*, (7.21) *holds if* $p \leq \frac{1}{2}$ *and* (7.22) *holds if* $p = \frac{1}{2}$. *If the policy is bounded* (*and* F_τ *has finite range*), *then* $E[F_\tau] \leq F_0$ *for* $p \leq \frac{1}{2}$ *and* $E[F_\tau] = F_0$ *for* $p = \frac{1}{2}$.

Example 7.6. The gambler has initial capital a and plays at unit stakes until his capital increases to c ($0 \leq a \leq c$) or he is ruined. Here $F_0 = a$ and $W_n = 1$, and so $F_n = a + S_n$. The policy is bounded by c, and F_τ is c or 0 according as the gambler succeeds or fails. If $p = \frac{1}{2}$ and if s is the probability of success, then $a = F_0 = E[F_\tau] = sc$. Thus $s = a/c$. This gives a new derivation of (7.7) for the case $p = \frac{1}{2}$. The argument assumes however that play is certain to terminate. If $p \leq \frac{1}{2}$, Theorem 7.2 only gives $s \leq a/c$, which is weaker than (7.7). ■

Example 7.7. Suppose as before that $F_0 = a$ and $W_n = 1$, so that $F_n = a + S_n$, but suppose the stopping rule is to quit as soon as F_n reaches $a + b$. Here F_n^* is bounded above by $a + b$ but is not bounded below. If $p = \frac{1}{2}$, the gambler is by (7.8) certain to achieve his goal, so that $F_\tau = a + b$. In this case $F_0 = a < a + b = E[F_\tau]$. This illustrates the effect of infinite capital. It also illustrates the need for uniform boundedness in Theorem 5.4 (compare Example 5.7). ■

For some other systems (gamblers call them "martingales"), see the problems. For most such systems there is a large chance of a small gain and a small chance of a large loss.

Bold Play*

The formula (7.7) gives the chance that a gambler betting unit stakes can increase his fortune from a to c before being ruined. Suppose that a and c happen to be even and that at each trial the wager is two units instead of

[†]See Problem 7.11.
*This topic may be omitted.

one. Since this has the effect of halving a and c, the chance of success is now

$$\frac{\rho^{a/2} - 1}{\rho^{c/2} - 1} = \frac{\rho^a - 1}{\rho^c - 1} \frac{\rho^{c/2} + 1}{\rho^{a/2} + 1}, \qquad \frac{q}{p} = \rho \neq 1.$$

If $\rho > 1$ $(p < \frac{1}{2})$, the second factor on the right exceeds 1: Doubling the stakes increases the probability of success in the unfavorable case $\rho > 1$. In the case $\rho = 1$, the probability remains the same.

There is a sense in which large stakes are optimal. It will be convenient to rescale so that the initial fortune satisfies $0 \le F_0 \le 1$ and the goal is 1. The policy of *bold play* is this: At each stage the gambler bets his entire fortune, unless a win would carry him past his goal of 1, in which case he bets just enough that a win would exactly achieve that goal:

$$(7.25) \qquad W_n = \begin{cases} F_{n-1} & \text{if } 0 \le F_{n-1} \le \frac{1}{2}, \\ 1 - F_{n-1} & \text{if } \frac{1}{2} \le F_{n-1} \le 1. \end{cases}$$

(It is convenient to allow even irrational fortunes.) As for stopping, the policy is to quit as soon as F_n reaches 0 or 1.

Suppose that play has not terminated by time $k - 1$; under the policy (7.25), if play is not to terminate at time k, then X_k must be $+1$ or -1 according as $F_{k-1} \le \frac{1}{2}$ or $F_{k-1} \ge \frac{1}{2}$, and the conditional probability of this is at most $m = \max\{p, q\}$. It follows by induction that the probability that bold play continues beyond time n is at most m^n, and so play is certain to terminate (τ is finite with probability 1).

It will be shown that in the subfair case, bold play maximizes the probability of successfully reaching the goal of 1. This is the *Dubins–Savage theorem*. It will further be shown that there are other policies that are also optimal in this sense, and this maximum probability will be calculated. Bold play can be substantially better than betting at constant stakes. This contrasts with Theorems 7.1 and 7.2 concerning respects in which gambling systems are worthless.

From now on, consider only policies π that are bounded by 1 (see (7.24)). Suppose further that play stops as soon as F_n reaches 0 or 1 and that this is certain eventually to happen. Since F_τ assumes the values 0 and 1, and since $[F_\tau = x] = \bigcup_{n=0}^{\infty} [\tau = n] \cap [F_n = x]$ for $x = 0$ and $x = 1$, F_τ is a simple random variable. Bold play is one such policy π.

The policy π leads to success if $F_\tau = 1$. Let $Q_\pi(x)$ be the probability of this for an initial fortune $F_0 = x$:

$$(7.26) \qquad Q_\pi(x) = P[F_\tau = 1] \quad \cdot \quad \text{for } F_0 = x.$$

Since F_n is a function $\psi_n(F_0, X_1(\omega), \ldots, X_n(\omega)) = \Psi_n(F_0, \omega)$, (7.26) in expanded notation is $Q_\pi(x) = P[\omega: \Psi_{\tau(x,\omega)}(x, \omega) = 1]$. As π specifies that play stops at the boundaries 0 and 1,

$$
(7.27) \qquad
\begin{aligned}
&Q_\pi(0) = 0, \quad Q_\pi(1) = 1, \\
&0 \le Q_\pi(x) \le 1, \qquad 0 \le x \le 1.
\end{aligned}
$$

Let Q be the Q_π for bold play. (The notation does not show the dependence of Q and Q_π on p, which is fixed.)

Theorem 7.3. *In the subfair case,* $Q_\pi(x) \le Q(x)$ *for all* π *and all* x.

PROOF. Under the assumption $p \le q$, it will be shown later that

$$(7.28) \qquad Q(x) \ge pQ(x + t) + qQ(x - t), \qquad 0 \le x - t \le x \le x + t \le 1.$$

This can be interpreted as saying that the chance of success under bold play starting at x is at least as great as the chance of success if the amount t is wagered and bold play then pursued from $x + t$ in case of a win and from $x - t$ in case of a loss. Under the assumption of (7.28), optimality can be proved as follows.

Consider a policy π, and let F_n and F_n^* be the simple random variables defined by (7.14) and (7.19) *for this policy.* Now $Q(x)$ is a real function, and so $Q(F_n^*)$ is also a simple random variable; it can be interpreted as the conditional chance of success if π is replaced by bold play after time n. By (7.20), $F_n^* = x + tX_n$ if $F_{n-1}^* = x$ and $W_n^* = t$. Therefore,

$$
Q(F_n^*) = \sum_{x, t} I_{[F_{n-1}^* = x, \, W_n^* = t]} Q(x + tX_n),
$$

where x and t vary over the (finite) ranges of F_{n-1}^* and W_n^*, respectively.

For each x and t, the indicator above is measurable $\mathcal{F}_{n-1}$ and $Q(x + tX_n)$ is measurable $\sigma(X_n)$; since the X_n are independent, (5.25) and (5.17) give

$$(7.29) \qquad E[Q(F_n^*)] = \sum_{x, t} P[F_{n-1}^* = x, \, W_n^* = t] E[Q(x + tX_n)]$$

By (7.28), $E[Q(x + tX_n)] \le Q(x)$ if $0 \le x - t \le x \le x + t \le 1$. As it is assumed of π that F_n^* lies in $[0,1]$ (that is, $W_n^* \le \min\{F_{n-1}^*, 1 - F_{n-1}^*\}$), the probability

in (7.29) is 0 unless x and t satisfy this constraint. Therefore,

$$E[Q(F_n^*)] \leq \sum_{x,t} P[F_{n-1}^* = x, W_n^* = t]Q(x)$$

$$= \sum_x P[F_{n-1}^* = x]Q(x) = E[Q(F_{n-1}^*)].$$

This is true for each n, and so $E[Q(F_n^*)] \leq E[Q(F_0^*)] = Q(F_0)$. Since $Q(F_n^*)$ $= Q(F_\tau)$ for $n \geq \tau$, Theorem 5.4 implies that $E[Q(F_\tau)] \leq Q(F_0)$. Since $x = 1$ implies that $Q(x) = 1$, $P[F_\tau = 1] \leq E[Q(F_\tau)] \leq Q(F_0)$. Thus $Q_\pi(F_0) \leq Q(F_0)$ for the policy π, whatever F_0 may be.

It remains to analyze Q and prove (7.28). Everything hinges on the functional equation

$$(7.30) \qquad Q(x) = \begin{cases} pQ(2x), & 0 \leq x \leq \frac{1}{2}, \\ p + qQ(2x - 1), & \frac{1}{2} \leq x \leq 1. \end{cases}$$

For $x = 0$ and $x = 1$ this is obvious because $Q(0) = 0$ and $Q(1) = 1$. The idea is this: Suppose that the initial fortune is x. If $x \leq \frac{1}{2}$, the first stake under bold play is x; if the gambler is to succeed in reaching 1, he must win the first trial (probability p) and then from his new fortune $x + x = 2x$ go on to succeed (probability $Q(2x)$); this makes the first half of (7.30) plausible. If $x \geq \frac{1}{2}$, the first stake is $1 - x$; the gambler can succeed either by winning the first trial (probability p) or by losing the first trial (probability q) and then going on from his new fortune $x - (1 - x) = 2x - 1$ to succeed (probability $Q(2x - 1)$); this makes the second half of (7.30) plausible.

It is also intuitively clear that $Q(x)$ must be an increasing function of x $(0 \leq x \leq 1)$: the more money the gambler starts with, the better off he is. Finally, it is intuitively clear that $Q(x)$ ought to be a continuous function of the initial fortune x.

A formal proof of (7.30) can be constructed as for the difference equation (7.5). If $\beta(x)$ is x for $x \leq \frac{1}{2}$ and $1 - x$ for $x \geq \frac{1}{2}$, then under bold play $W_n = \beta(F_{n-1})$. Starting from $f_0(x) = x$, recursively define

$$f_n(x; x_1, \ldots, x_n) = f_{n-1}(x; x_1, \ldots, x_{n-1}) + \beta(f_{n-1}(x; x_1, \ldots, x_{n-1}))x_n.$$

Then $F_n = f_n(F_0; X_1, \ldots, X_n)$. Now define

$$g_n(x; x_1, \ldots, x_n) = \max_{0 \leq k \leq n} f_k(x; x_1, \ldots, x_k).$$

If $F_0 = x$, then $T_n(x) = [g_n(x; X_1, \ldots, X_n) = 1]$ is the event that bold play will by time n successfully increase the gambler's fortune to 1. From the recursive definition it

follows by induction on n that for $n \geq 1$, $f_n(x; x_1, \ldots, x_n) = f_{n-1}(x + \beta(x)x_1;$ $x_2, \ldots, x_n)$ and hence that $g_n(x; x_1, \ldots, x_n) = \max\{x, g_{n-1}(x + \beta(x)x_1; x_2, \ldots, x_n)\}$. Since $x = 1$ implies $g_{n-1}(x + \beta(x)x_1; x_2, \ldots, x_n) \geq x + \beta(x)x_1 = 1$, $T_n(x) = [g_{n-1}(x + \beta(x)X_1; X_2, \ldots, X_n) = 1]$, and since the X_i are independent and identically distributed, $P(T_n(x)) = P([X_1 = +1] \cap T_n(x)) + P([X_1 = -1] \cap T_n(x)) = pP[g_{n-1}(x + \beta(x); X_2, \ldots, X_n) = 1] + qP[g_{n-1}(x - \beta(x); X_2, \ldots, X_n) = 1] = pP(T_{n-1}(x + \beta(x))) + qP(T_{n-1}(x - \beta(x)))$. Letting $n \to \infty$ now gives $Q(x) = pQ(x + \beta(x)) + qQ(x - \beta(x))$, which reduces to (7.30) because $Q(0) = 0$ and $Q(1) = 1$.

Suppose that $y = f_{n-1}(x; x_1, \ldots, x_{n-1})$ is nondecreasing in x. If $x_n = +1$, then $f_n(x; x_1, \ldots, x_n)$ is $2y$ if $0 \leq y \leq \frac{1}{2}$ and 1 if $\frac{1}{2} \leq y \leq 1$; if $x_n = -1$, then $f_n(x; x_1, \ldots, x_n)$ is 0 if $0 \leq y \leq \frac{1}{2}$ and $2y - 1$ if $\frac{1}{2} \leq y \leq 1$. In any case, $f_n(x; x_1, \ldots, x_n)$ is also nondecreasing in x, and by induction this is true for every n. It follows that the same is true of $g_n(x; x_1, \ldots, x_n)$, of $P(T_n(x))$, and of $Q(x)$. Thus $Q(x)$ is nondecreasing.

Since $Q(1) = 1$, (7.30) implies that $Q(\frac{1}{2}) = pQ(1) = p$, $Q(\frac{1}{4}) = pQ(\frac{1}{2}) = p^2$, $Q(\frac{3}{4}) = p + qQ(\frac{1}{2}) = p + pq$ More generally, if $p_0 = p$ and $p_1 = q$, then

$$(7.31) \qquad Q\left(\frac{k}{2^n}\right) = \sum \left[p_{u_1} \cdots p_{u_n} : \sum_{i=1}^{r} \frac{u_i}{2^i} < \frac{k}{2^n} \right], \qquad 0 < k \leq 2^n, \quad n \geq 1,$$

the sum extending over n-tuples $(u_1, \ldots, u_n)$ of 0's and 1's satisfying the condition indicated. Indeed, it is easy to see that (7.31) is the same thing as

$$(7.32) \qquad Q(.u_1 \ldots u_n + 2^{-n}) - Q(.u_1 \ldots u_n) = p_{u_1} p_{u_2} \cdots p_{u_n}$$

for each dyadic rational $.u_1 \ldots u_n$ of rank n. If $.u_1 \ldots u_n + 2^{-n} \leq \frac{1}{2}$, then $u_1 = 0$ and by (7.30) the difference in (7.32) is $p_0[Q(.u_2 \ldots u_n + 2^{-n+1}) - Q(.u_2 \ldots u_n)]$. But (7.32) follows inductively from this and a similar relation for the case $.u_1 \ldots u_n \geq \frac{1}{2}$.

Therefore $Q(k2^{-n}) - Q((k-1)2^{-n})$ is bounded by $\max\{p^n, q^n\}$, and so by monotonicity Q is continuous. Since (7.32) is positive, it follows that Q is strictly increasing over $[0, 1]$.

Thus Q is continuous and increasing and satisfies (7.30). The inequality (7.28) is still to be proved. It is equivalent to the assertion that

$$\Delta(r, s) = Q(a) - pQ(s) - qQ(r) \geq 0$$

if $0 \leq r \leq s \leq 1$, where a stands for the average: $a = \frac{1}{2}(r + s)$. Since Q is continuous, it suffices to prove the inequality for r and s of the form $k/2^n$, and this will be done by induction on n. Checking all cases disposes of $n = 0$. Assume that the inequality holds for a particular n, and that r and s have the form $k/2^{n+1}$. There are four cases to consider.

CASE 1. $s \leq \frac{1}{2}$. By the first part of (7.30), $\Delta(r, s) = p\Delta(2r, 2s)$. Since $2r$ and $2s$ have the form $k/2^n$, the induction hypothesis implies that $\Delta(2r, 2s) \geq 0$.

CASE 2. $\frac{1}{2} \le r$. By the second part of (7.30),

$$\Delta(r, s) = q\Delta(2r - 1, 2s - 1) \ge 0.$$

CASE 3. $r \le a \le \frac{1}{2} \le s$. By (7.30),

$$\Delta(r, s) = pQ(2a) - p[p + qQ(2s - 1)] - q[pQ(2r)].$$

From $\frac{1}{2} \le s \le r + s = 2a \le 1$, follows $Q(2a) = p + qQ(4a - 1)$; and from $0 \le 2a - \frac{1}{2} \le \frac{1}{2}$, follows $Q(2a - \frac{1}{2}) = pQ(4a - 1)$. Therefore, $pQ(2a) = p^2 + qQ(2a - \frac{1}{2})$, and it follows that

$$\Delta(r, s) = q[Q(2a - \frac{1}{2}) - pQ(2s - 1) - pQ(2r)].$$

Since $p \le q$, the right side does not increase if either of the two p's is changed to q. Hence

$$\Delta(r, s) \ge q \max[\Delta(2r, 2s - 1), \Delta(2s - 1, 2r)].$$

The induction hypothesis applies to $2r \le 2s - 1$ or to $2s - 1 \le 2r$, as the case may be, so one of the two Δ's on the right is nonnegative.

CASE 4. $r \le \frac{1}{2} \le a \le s$. By (7.30),

$$\Delta(r, s) = pq + qQ(2a - 1) - pqQ(2s - 1) - pqQ(2r).$$

From $0 \le 2a - 1 = r + s - 1 \le \frac{1}{2}$, follows $Q(2a - 1) = pQ(4a - 2)$; and from $\frac{1}{2} \le 2a - \frac{1}{2} = r + s - \frac{1}{2} \le 1$, follows $Q(2a - \frac{1}{2}) = p + qQ(4a - 2)$. Therefore, $qQ(2a - 1) = pQ(2a - \frac{1}{2}) - p^2$, and it follows that

$$\Delta(r, s) = p[q - p + Q(2a - \frac{1}{2}) - qQ(2s - 1) - qQ(2r)].$$

If $2s - 1 \le 2r$, the right side here is

$$p[(q - p)(1 - Q(2r)) + \Delta(2s - 1, 2r)] \ge 0.$$

If $2r \le 2s - 1$, the right side is

$$p[(q - p)(1 - Q(2s - 1)) + \Delta(2r, 2s - 1)] \ge 0.$$

This completes the proof of (7.28) and hence of Theorem 7.3. ∎

The equation (7.31) has an interesting interpretation. Let $Z_1, Z_2, \ldots$ be independent random variables satisfying $P[Z_n = 0] = p_0 = p$ and $P[Z_n = 1] = p_1 = q$. From $P[Z_n = 1 \text{ i.o.}] = 1$ and $\sum_{i > n} Z_i 2^{-i} \le 2^{-n}$ it follows that $P[\sum_{i=1}^{\infty} Z_i 2^{-i} \le k 2^{-n}] \le P[\sum_{i=1}^{n} Z_i 2^{-i} < k 2^{-n}] \le P[\sum_{i=1}^{\infty} Z_i 2^{-i} \le k 2^{-n}]$. Since by (7.31) the middle term is $Q(k 2^{-n})$,

$$(7.33) \qquad Q(x) = P\left[\sum_{i=1}^{\infty} Z_i 2^{-i} \le x \right]$$

holds for dyadic rational x and hence by continuity holds for all x. In Section 31, Q will reappear as a continuous, strictly increasing function singular in the sense of Lebesgue. On p. 408 is a graph for the case $p_0 = .25$.

Note that $Q(x) \equiv x$ in the fair case $p = \frac{1}{2}$. In fact, for a bounded policy Theorem 7.2 implies that $E[F_\tau] = F_0$ in the fair case, and if the policy is to stop as soon as the fortune reaches 0 or 1, then the chance of successfully reaching 1 is $P[F_\tau = 1] = E[F_\tau] = F_0$. Thus in the fair case with initial fortune x, the chance of success is x for *every* policy that stops at the boundaries, and x is an upper bound even if stopping earlier is allowed.

Example 7.8. The gambler of Example 7.1 has capital \$900 and goal \$1000. For a fair game $(p = \frac{1}{2})$ his chance of success is .9 whether he bets unit stakes or adopts bold play. At red-and-black $(p = \frac{18}{38})$, his chance of success with unit stakes is .00003; an approximate calculation based on (7.31) shows that under bold play his chance $Q(.9)$ of success increases to about .88, which compares well with the fair case. ∎

Example 7.9. In Example 7.2 the capital is \$100 and the goal \$20,000. At unit stakes the chance of successes is .005 for $p = \frac{1}{2}$ and 3×10^{-911} for $p = \frac{18}{38}$. Another approximate calculation shows that bold play at red-and-black gives the gambler probability about .003 of success, which again compares well with the fair case.

This example illustrates the point of Theorem 7.3. The gambler enters the casino knowing that he must by dawn convert his \$100 into \$20,000 or face certain death at the hands of criminals to whom he owes that amount. Only red-and-black is available to him. The question is not whether to gamble—he *must* gamble. The question is how to gamble so as to maximize the chance of survival, and bold play is the answer. ∎

There are policies other than the bold one that achieve the maximum success probability $Q(x)$. Suppose that as long as the gambler's fortune x is less than $\frac{1}{2}$ he bets x for $x \le \frac{1}{4}$ and $\frac{1}{2} - x$ for $\frac{1}{4} \le x \le \frac{1}{2}$. This is, in effect, the

bold-play strategy scaled down to the interval $[0, \frac{1}{2}]$, and so the chance he ever reaches $\frac{1}{2}$ is $Q(2x)$ for an initial fortune of x. Suppose further that if he does reach the goal of $\frac{1}{2}$, or if he starts with fortune at least $\frac{1}{2}$ in the first place, then he continues, but with ordinary bold play. For an initial fortune $x \geq \frac{1}{2}$, the overall chance of success is of course $Q(x)$, and for an initial fortune $x < \frac{1}{2}$, it is $Q(2x)Q(\frac{1}{2}) = pQ(2x) = Q(x)$. The success probability is indeed $Q(x)$ as for bold play, although the policy is different. With this example in mind, one can generate a whole series of distinct optimal policies.

Timid Play *

The optimality of bold play seems reasonable when one considers the effect of its opposite, timid play. Let the ϵ-*timid* policy be to bet $W_n = \min\{\epsilon, F_{n-1}, 1 - F_{n-1}\}$ and stop when F_n reaches 0 or 1. Suppose that $p < q$, fix an initial fortune $x = F_0$ with $0 \leq x < 1$, and consider what happens as $\epsilon \to 0$. By the strong law of large numbers, $\lim_n n^{-1} S_n = E[X_1] = p - q < 0$. There is therefore probability 1 that $\sup_k S_k < \infty$ and $\lim_n S_n = -\infty$. Given $\eta > 0$, choose ϵ so that $P[\sup_k(x + \epsilon S_k) < 1] > 1 - \eta$. Since $P(\bigcup_{n=1}^{\infty}[x + \epsilon S_n < 0]) = 1$, with probability at least $1 - \eta$ there exists an n such that $x + \epsilon S_n < 0$ and $\max_{k < n}(x + \epsilon S_k) < 1$. But under the ϵ-timid policy the gambler is in this circumstance ruined. If $Q_\epsilon(x)$ is the probability of success under the ϵ-timid policy, then $\lim_{\epsilon \to 0} Q_\epsilon(x) = 0$ for $0 \leq x < 1$. The law of large numbers carries the timid player to his ruin.[†]

PROBLEMS

7.1. A gambler with initial capital a plays until his fortune increases b units or he is ruined. Suppose that $\rho > 1$. The chance of success is multiplied by $1 + \theta$ if his initial capital is infinite instead of a. Show that $0 < \theta < (\rho^a - 1)^{-1} < (a(\rho - 1))^{-1}$; relate to Example 7.3.

7.2 As shown on p. 94, there is probability 1 that the gambler either achieves his goal of c or is ruined. For $p \neq q$, deduce this directly from the strong law of large numbers. Deduce it (for all p) via the Borel–Cantelli lemma from the fact that if play never terminates, there can never occur c successive $+1$'s.

7.3. 6 12↑   If V_n is the set of n-long sequences of ± 1's, the function b_n in (7.9) maps V_{n-1} into $\{0, 1\}$. A selection system is a sequence of such maps. Although there are uncountably many selection systems, how many have an *effective*

---

*This topic may be omitted
[†]For each ϵ, however, there exist optimal policies under which the bet never exceeds ϵ; see DUBINS & SAVAGE.

description in the sense of an algorithm or finite set of instructions by means of which a deputy (perhaps a machine) could operate the system for the gambler? An analysis of the question is a matter for mathematical logic, but one can see that there can be only countably many algorithms or finite sets of rules expressed in finite alphabets.

Let $Y_1^{(\sigma)}, Y_2^{(\sigma)}, \ldots$ be the random variables of Theorem 7.1 for a particular system σ, and let C_σ be the ω-set where every k-tuple of ± 1's (k arbitrary) occurs in $Y_1^{(\sigma)}(\omega), Y_2^{(\sigma)}(\omega), \ldots$ with the right asymptotic relative frequency (in the sense of Problem 6.12). Let C be the intersection of C_σ over all effective selection systems σ. Show that C lies in $\mathscr{F}$ (the σ-field in the probability space $(\Omega, \mathscr{F}, P)$ on which the X_n are defined) and that $P(C) = 1$. A sequence $(X_1(\omega), X_2(\omega), \ldots)$ for ω in C is called a *collective*: a subsequence chosen by any of the effective rules σ contains all k-tuples in the correct proportions

7.4. Let D_n be 1 or 0 according as $X_{2n-1} \neq X_{2n}$ or not, and let M_k be the time of the kth 1—the smallest n such that $\sum_{i=1}^{n} D_i = k$. Let $Z_k = X_{2M_k}$. In other words, look at successive nonoverlapping pairs (X_{2n-1}, X_{2n}), discard accordant $(X_{2n-1} = X_{2n})$ pairs, and keep the second element of discordant $(X_{2n-1} \neq X_{2n})$ pairs. Show that this process simulates a fair coin: $Z_1, Z_2, \ldots$ are independent and identically distributed and $P[Z_k = +1] = P[Z_k = -1] = \frac{1}{2}$, whatever p may be. Follow the proof of Theorem 7.1.

7.5. Suppose that a gambler with initial fortune 1 stakes a proportion θ ($0 < \theta < 1$) of his current fortune: $F_0 = 1$ and $W_n = \theta F_{n-1}$. Show that $F_n = \Pi_{k=1}^{n}(1 + \theta X_k)$ and hence that

$$\log F_n = \frac{n}{2}\left[\frac{S_n}{n} \log \frac{1+\theta}{1-\theta} + \log(1 - \theta^2) \right].$$

Show that $F_n \to 0$ with probability 1 in the subfair case.

7.6. In "doubling," $W_1 = 1$, $W_n = 2W_{n-1}$, and the rule is to stop after the first win. For any positive p, play is certain to terminate. Here $F_\tau = F_0 + 1$, but of course infinite capital is required. If $F_0 = 2^k - 1$ and W_n cannot exceed F_{n-1}, the probability of $F_\tau = F_0 + 1$ in the fair case is $1 - 2^{-k}$. Prove this via Theorem 7.2 and also directly.

7.7. In "progress and pinch," the wager, initially some integer, is increased by 1 after a loss and decreased by 1 after a win, the stopping rule being to quit if the next bet is 0. Show that play is certain to terminate if and only if $p \geq \frac{1}{2}$. Show that $F_\tau = F_0 + \frac{1}{2}W_1^2 + \frac{1}{2}(\tau - 1)$. Infinite capital is required.

7.8. Here is a common martingale. Just before the nth spin of the wheel, the gambler has before him a pattern $x_1, \ldots, x_k$ of positive numbers (k varies with n). He bets $x_1 + x_k$, or x_1 in case $k = 1$. If he loses, at the next stage he uses the pattern $x_1, \ldots, x_k, x_1 + x_k$ (x_1, x_1 in case $k = 1$). If he wins, at the next stage he uses the pattern $x_2, \ldots, x_{k-1}$, unless k is 1 or 2, in which case he quits. Show

that play is certain to terminate if $p > \frac{1}{3}$ and that the ultimate gain is the sum of the numbers in the initial pattern. Infinite capital is again required.

7.9. Suppose that $W_k = 1$, so that $F_k = F_0 + S_k$. Suppose that $p \geq q$ and τ is a stopping time such that $1 \leq \tau \leq n$ with probability 1. Show that $E[F_\tau] \leq E[F_n]$, with equality in case $p = q$. Interpret this result in terms of a stock option that must be exercised by time n, where $F_0 + S_k$ represents the price of the stock at time k.

7.10. For a given policy, let A_n^* be the fortune of the gambler's adversary at time n. Consider these conditions on the policy. (i) $W_n^* \leq F_{n-1}^*$; (ii) $W_n^* \leq A_{n-1}^*$; (iii) $F_n^* + A_n^*$ is constant. Interpret each condition, and show that together they imply that the policy is bounded in the sense of (7.24).

7.11. Show that F_τ has infinite range if $F_0 = 1$, $W_n = 2^{-n}$, and τ is the smallest n for which $X_n = +1$.

7.12. Let u be a real function on $[0, 1]$, $u(x)$ representing the *utility* of the fortune x. Consider policies bounded by 1; see (7.24). Let $Q_\pi(F_0) = E[u(F_\tau)]$; this represents the expected utility under the policy π of an initial fortune F_0. Suppose of a policy π_0 that

$$(7.34) \qquad u(x) \leq Q_{\pi_0}(x), \qquad 0 \leq x \leq 1,$$

and that

$$(7.35) \qquad Q_{\pi_0}(x) \geq p Q_{\pi_0}(x + t) + q Q_{\pi_0}(x - t),$$

$$0 \leq x - t \leq x \leq x + t \leq 1.$$

Show that $Q_\pi(x) \leq Q_{\pi_0}(x)$ for all x and all policies π. Such a π_0 is optimal.

Theorem 7.3 is the special case of this result for $p \leq \frac{1}{2}$, bold play in the role of π_0, and $u(x) = 1$ or $u(x) = 0$ according as $x = 1$ or $x < 1$.

The condition (7.34) says that gambling with policy π_0 is at least as good as not gambling at all; (7.35) says that, although the prospects even under π_0 become on the average less sanguine as time passes, it is better to use π_0 now than to use some other policy for one step and then change to π_0.

7.13. The functional equation (7.30) and the assumption that Q is bounded suffice to determine Q completely. First, $Q(0)$ and $Q(1)$ must be 0 and 1, respectively, and so (7.31) holds. Let $T_0 x = \frac{1}{2}x$ and $T_1 x = \frac{1}{2}x + \frac{1}{2}$; let $f_0 x = px$ and $f_1 x = p + qx$. Then $Q(T_{u_1} \cdots T_{u_n} x) = f_{u_1} \cdots f_{u_n} Q(x)$. If the binary expansions of x and y both begin with the digits $u_1, \ldots, u_n$, they have the form $x = T_{u_1} \ldots T_{u_n} x'$ and $y = T_{u_1} \cdots T_{u_n} y'$. If K bounds Q and if $m = \max\{p, q\}$, it follows that $|Q(x) - Q(y)| \leq K m^n$. Therefore, Q is continuous and satisfies (7.31) and (7.33).

SECTION 8. MARKOV CHAINS

As Markov chains illustrate in a clear and striking way the connection between probability and measure, their basic properties are developed here in a measure-theoretic setting.

Definitions

Let S be a finite or countable set. Suppose that to each pair i and j in S there is assigned a nonnegative number p_{ij} and that these numbers satisfy the constraint

$$(8.1) \qquad \sum_{j \in S} p_{ij} = 1, \qquad i \in S.$$

Let $X_0, X_1, X_2, \ldots$ be a sequence of random variables whose ranges are contained in S. The sequence is a *Markov chain* or *Markov process* if

$$(8.2) \qquad P\big[X_{n+1} = j | X_0 = i_0, \ldots, X_n = i_n \big]$$
$$= P\big[X_{n+1} = j | X_n = i_n \big] = p_{i_n j}$$

for every n and every sequence $i_0, \ldots, i_n$ in S for which $P[X_0 = i_0, \ldots, X_n = i_n] > 0$. The set S is the *state space* or *phase space* of the process, and the p_{ij} are the *transition probabilities*. Part of the defining condition (8.2) is that the transition probability

$$(8.3) \qquad P\big[X_{n+1} = j | X_n = i \big] = p_{ij}$$

does not vary with n.[†]

The elements of S are thought of as the possible states of a *system*, X_n representing the state at *time n*. The sequence or process $X_0, X_1, X_2, \ldots$ then represents the history of the system, which evolves in accordance with the probability law (8.2). The conditional distribution of the *next* state X_{n+1} given the *present* state X_n must not further depend on the *past* $X_0, \ldots, X_{n-1}$. This is what (8.2) requires, and it leads to a copious theory.

The *initial probabilities* are

$$(8.4) \qquad \alpha_i = P[X_0 = i].$$

The α_i are nonnegative and add to 1, but the definition of Markov chain places no further restrictions on them.

[†]Sometimes in the definition of the Markov chain $P[X_{n+1} = j | X_n = i]$ is allowed to depend on n. A chain satisfying (8.3) is then said to have *stationary transition probabilities*, a phrase that will be omitted here because (8.3) will always be assumed.

The following examples illustrate some of the possibilities. In each one, the state space S and the transition probabilities p_{ij} are described, but the underlying probability space $(\Omega, \mathscr{F}, P)$ and the X_n are left unspecified for now: see Theorem 8.1.[†]

Example 8.1. The Bernoulli–Laplace model of diffusion. Imagine r black balls and r white balls distributed between two boxes, with the constraint that each box contains r balls. The state of the system is specified by the number of white balls in the first box, so that the state space is $S = \{0, 1, \ldots, r\}$. The transition mechanism is this: at each stage one ball is chosen at random from each box and the two are interchanged. If the present state is i, the chance of a transition to $i - 1$ is the chance i/r of drawing one of the i white balls from the first box times the chance i/r of drawing one of the i black balls from the second box. Together with similar arguments for the other possibilities, this shows that the transition probabilities are

$$p_{i,i-1} = \left(\frac{i}{r}\right)^2, \qquad p_{i,i+1} = \left(\frac{r-i}{r}\right)^2, \qquad p_{ii} = 2\frac{i(r-i)}{r^2},$$

the others being 0. This is the probablistic analogue of the model for the flow of two liquids between two containers.   ∎

The p_{ij} form the *transition matrix* $P = [p_{ij}]$ of the process. A *stochastic matrix* is one whose entries are nonnegative and satisfy (8.1); the transition matrix of course has this property.

Example 8.2. Random walk with absorbing barriers. Suppose that $S = \{0, 1, \ldots, r\}$ and

$$P = \begin{bmatrix} 1 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 \\ q & 0 & p & 0 & \ldots & 0 & 0 & 0 & 0 \\ 0 & q & 0 & p & \ldots & 0 & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & 0 & \ldots & q & 0 & p & 0 \\ 0 & 0 & 0 & 0 & \ldots & 0 & q & 0 & p \\ 0 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 & 1 \end{bmatrix}$$

That is, $p_{i,i+1} = p$ and $p_{i,i-1} = q = 1 - p$ for $0 < i < r$ and $p_{00} = p_{rr} = 1$. The chain represents a particle in *random walk*. The particle moves one unit to the right or left, the respective probabilities being p and q, except that each of 0 and r is an *absorbing* state—once the particle enters, it cannot leave. The state can also be viewed as a gambler's fortune; absorption in 0

represents ruin for the gambler, absorption in r ruin for his adversary (see Section 7). The gambler's initial fortune is usually regarded as nonrandom, so that (see (8.4)) $\alpha_i = 1$ for some i.   ■

Example 8.3.   Unrestricted random walk. Let S consist of all the integers $i = 0, \pm 1, \pm 2, \ldots$, and take $p_{i, i+1} = p$ and $p_{i, i-1} = q = 1 - p$. This chain represents a random walk without barriers, the particle being free to move anywhere on the integer lattice. The walk is *symmetric* if $p = q$.   ■

The state space may, as in the preceding example, be countably infinite. If so, the Markov chain consists of functions X_n on a probability space $(\Omega, \mathcal{F}, P)$, but these will have infinite range and hence will not be random variables in the sense of the preceding sections. This will cause no difficulty, however, because expected values of the X_n will not be considered. All that is required is that for each $i \in S$ the set $\{\omega: X_n(\omega) = i\}$ lie in $\mathcal{F}$ and hence have a probability.

Example 8.4.   Symmetric random walk in space. Let S consist of the integer lattice points in k-dimensional Euclidean space R^k; $x = (x_1, \ldots, x_k)$ lies in S if the coordinates are all integers. Now x has $2k$ neighbors, points of the form $y = (x_1, \ldots, x_i \pm 1, \ldots, x_k)$; for each such y let $p_{xy} = (2k)^{-1}$. The chain represents a particle moving randomly in space; for $k = 1$ it reduces to Example 8.3 with $p = q = \frac{1}{2}$. The cases $k \le 2$ and $k \ge 3$ exhibit an interesting difference. If $k \le 2$, the particle is certain to return to its initial position, but this is not so if $k \ge 3$; see Example 8.6.   ■

Since the state space in this example is not a subset of the line, the $X_0, X_1, \ldots$ do not assume real values. This is immaterial because expected values of the X_n play no role. All that is necessary is that X_n be a mapping from Ω into S (finite or countable) such that $\{\omega: X_n(\omega) = i\} \in \mathcal{F}$ for $i \in S$. There will be expected values $E[f(X_n)]$ for real functions f on S with finite range, but then $f(X_n(\omega))$ is a simple random variable as defined before.

Example 8.5.   A selection problem. A princess must chose from among r suitors. She is definite in her preferences and if presented with all r at once could choose her favorite and could even rank the whole group. They are ushered into her presence one by one in random order, however, and she must at each stage either stop and accept the suitor or else reject him and proceed in the hope that a better one will come along. What strategy will maximize her chance of stopping with the best suitor of all?

Shorn of some details, the analysis is this. Let $S_1, S_2, \ldots, S_r$ be the suitors in order of presentation; this sequence is a random permutation of the set of suitors. Let $X_1 = 1$ and let $X_2, X_3, \ldots$ be the successive positions of suitors who dominate (are preferable to) all their predecessors. Thus $X_2 = 4$ and $X_3 = 6$ means that S_1 dominates S_2 and S_3 but S_4 dominates S_1, S_2, S_3, and that S_4 dominates S_5 but S_6 dominates $S_1, \ldots, S_5$. There can be at most r of these dominant suitors; if there are exactly m, $X_{m+1} = X_{m+2} = \cdots = r + 1$ by convention.

As the suitors arrive in random order, the chance that S_i ranks highest among $S_1, \ldots, S_i$ is $(i-1)!/i! = 1/i$. The chance that S_j ranks highest among $S_1, \ldots, S_j$ and S_i ranks next is $(j-2)!/j! = 1/j(j-1)$. This leads to a chain with transition probabilities[†]

$$(8.5) \qquad P[X_{n+1} = j | X_n = i] = \frac{i}{j(j-1)}, \qquad 1 \le i < j \le r.$$

If $X_n = i$, then $X_{n+1} = r+1$ means that S_i dominates $S_{i+1}, \ldots, S_r$ as well as $S_1, \ldots, S_i$, and the conditional probability of this is

$$(8.6) \qquad P[X_{n+1} = r+1 | X_n = i] = \frac{i}{r}, \qquad 1 \le i \le r$$

As downward transitions are impossible and $r+1$ is absorbing, this specifies a transition matrix for $S = \{1, 2, \ldots, r+1\}$.

It is quite clear that in maximizing her chance of selecting the best suitor of all, the princess should reject those who do not dominate their predecessors. Her strategy therefore will be to stop with the suitor in position X_τ, where τ is a random variable representing her strategy. Since her decision to stop must depend only on the suitors she has seen thus far, the event $[\tau = n]$ must lie in $\sigma(X, \ldots, X_n)$ If $X_\tau = i$, then by (8.6) the conditional probability of success is $f(i) = i/r$. The probability of success is therefore $E[f(X_\tau)]$, and the problem is to choose the strategy τ so as to maximize it. For the solution, see Example 8 17.[‡] ■

Higher-Order Transitions

The properties of the Markov chain are entirely determined by the transition and initial probabilities. The chain rule (4.2) for conditional probabilities gives

$$P[X_0 = i_0, X_1 = i_1, X_2 = i_2]$$
$$= P[X_0 = i_0] P[X_1 = i_1 | X_0 = i_0] P[X_2 = i_2 | X_0 = i_0, X_1 = i_1]$$
$$= \alpha_{i_0} p_{i_0 i_1} p_{i_1 i_2}.$$

Similarly,

$$(8.7) \qquad P[X_t = i_t, 0 \le t \le m] = \alpha_{i_0} p_{i_0 i_1} \cdots p_{i_{m-1} i_m}$$

for any sequence $i_0, i_1, \ldots, i_m$ of states.
    Further,

$$(8.8) \quad P[X_{m+t} = j_t, 1 \le t \le n | X_s = i_s, 0 \le s \le m] = p_{i_m j_1} p_{j_1 j_2} \cdots p_{j_{n-1} j_n},$$

[†]The details can be found in DYNKIN & YUSHKEVICH. Chapter III.
[‡]With the princess replaced by an executive and the suitors by applicants for an office job, this is known as the *secretary problem*

as follows by expressing the conditional probability as a ratio and applying (8.7) to numerator and denominator. Adding out the intermediate states now gives the formula

$$(8.9) \qquad p_{ij}^{(n)} = P[X_{m+n} = j | X_m = i]$$

$$= \sum_{k_1 \quad k_{n-1}} p_{ik_1} p_{k_1 k_2} \cdots p_{k_{n-1} j}$$

(the k_l range over S) for the *nth-order transition probabilities*.

Notice that $p_{ij}^{(n)}$ is the entry in position (i, j) of P^n, the nth power of the transition matrix P. If S is infinite, P is a matrix with infinitely many rows and columns; as the terms in (8.9) are nonnegative, there are no convergence problems. It is natural to put

$$p_{ij}^{(0)} = \delta_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

Then P^0 is the identity I, as it should be. From (8.1) and (8.9) follow

$$(8.10) \qquad p_{ij}^{(m+n)} = \sum_{\nu} p_{i\nu}^{(m)} p_{\nu j}^{(n)}, \qquad \sum_{j} p_{ij}^{(n)} = 1.$$

An Existence Theorem

Theorem 8.1. *Suppose that $P = [p_{ij}]$ is a stochastic matrix and that α_i are nonnegative numbers satisfying $\sum_{i \in S} \alpha_i = 1$. There exists on some $(\Omega, \mathscr{F}, P)$ a Markov chain $X_0, X_1, X_2, \ldots$ with initial probabilities α_i and transition probabilities p_{ij}.*

PROOF. Reconsider the proof of Theorem 5.3. There the space $(\Omega, \mathscr{F}, P)$ was the unit interval, and the central part of the argument was the construction of the decompositions (5.13). Suppose for the moment that $S = \{1, 2, \ldots\}$. First construct a partition $I_1^{(0)}, I_2^{(0)}, \ldots$ of $(0, 1]$ into countably many[†] subintervals of lengths (P is again Lebesgue measure) $P(I_i^{(0)}) = \alpha_i$. Next decompose each $I_i^{(0)}$ into subintervals $I_{ij}^{(1)}$ of lengths $P(I_{ij}^{(1)}) = \alpha_i p_{ij}$. Continuing inductively gives a sequence of partitions $\{I_{i_0 \cdots i_n}^{(n)}: i_0, \ldots, i_n = 1, 2, \ldots\}$ such that each refines the preceding and $P(I_{i_0 \cdots i_n}^{(n)}) = \alpha_{i_0} p_{i_0 i_1} \cdots p_{i_{n-1} i_n}$.

Put $X_n(\omega) = i$ if $\omega \in \bigcup_{i_0 \cdots i_{n-1}} I_{i_0 \cdots i_{n-1} i}^{(n)}$. It follows just as in the proof of Theorem 5.3 that the set $[X_0 = i_0, \ldots, X_n = i_n]$ coincides with the interval $I_{i_0 \cdots i_n}^{(n)}$. Thus $P[X_0 = i_0, \ldots, X_n = i_n] = \alpha_{i_0} p_{i_0 i_1} \cdots p_{i_{n-1} i_n}$. From this it follows immediately that (8.4) holds and that the first and third members of

---

[†]If $\delta_1 + \delta_2 + \cdots = b - a$ and $\delta_i \geq 0$, then $I_i = (b - \sum_{j \leq i} \delta_j, b - \sum_{j < i} \delta_j]$, $i = 1, 2, \ldots$, decompose $(a, b]$ into intervals of lengths δ_i.

(8.2) are the same. As for the middle member, it is $P[X_n = i_n, X_{n+1} = j]/P[X_n = i_n]$; the numerator is $\sum \alpha_{i_0} p_{i_0 i_1} \cdots p_{i_{n-1} i_n} p_{i_n j}$ the sum extending over all $i_0, \ldots, i_{n-1}$, and the denominator is the same thing without the factor $p_{i_n j}$, which means that the ratio is $p_{i_n j}$, as required.

That completes the construction for the case $S = \{1, 2, \ldots\}$. For the general countably infinite S, let g be a one-to-one mapping of $\{1, 2, \ldots\}$ onto S, and replace the X_n as already constructed by $g(X_n)$; the assumption $S = \{1, 2, \ldots\}$ was merely a notational convenience. The same argument obviously works if S is finite.[†]   ■

Although strictly speaking the Markov chain *is* the sequence X_0, $X_1, \ldots$, one often speaks as though the chain were the matrix P together with the initial probabilities α_i or even P with some unspecified set of α_i. Theorem 8.1 justifies this attitude: For given P and α_i the corresponding X_n do exist, and the apparatus of probability theory—the Borel–Cantelli lemmas and so on—is available for the study of P and of systems evolving in accordance with the Markov rule.

From now on fix a chain $X_0, X_1, \ldots$ satisfying $\alpha_i > 0$ for all i. Denote by P_i probabilities conditional on $[X_0 = i]$: $P_i(A) = P[A \mid X_0 = i]$. Thus

$$(8.11) \qquad P_i[X_t = i_t, 1 \le t \le n] = p_{i i_1} p_{i_1 i_2} \cdots p_{i_{n-1} i_n}$$

by (8.8). The interest centers on these conditional probabilities, and the actual initial probabilities α_i are now largely irrelevant.

From (8.11) follows

$$(8.12) \qquad P_i[X_1 = i_1, \ldots, X_m = i_m, X_{m+1} = j_1, \ldots, X_{m+n} = j_n]$$
$$= P_i[X_1 = i_1, \ldots, X_m = i_m] P_{i_m}[X_1 = j_1, \ldots, X_n = j_n].$$

Suppose that I is a set (finite or infinite) of m-long sequences of states, J is a set of n-long sequences of states, and every sequence in I ends in j. Adding both sides of (8.12) for $(i_1, \ldots, i_m)$ ranging over I and $(j_1, \ldots, j_n)$ ranging over J gives

$$(8.13) \qquad P_i[(X_1, \ldots, X_m) \in I, (X_{m+1}, \ldots, X_{m+n}) \in J]$$
$$= P_i[(X_1, \ldots, X_m) \in I] P_j[(X_1, \ldots, X_n) \in J].$$

For this to hold it is essential that each sequence in I end in j. The formulas (8.12) and (8.13) are of central importance.

[†]For a different approach in the finite case, see Problem 8.1.

Transience and Persistence

Let

$$(8.14) \qquad f_{ij}^{(n)} = P_i[X_1 \neq j, \ldots, X_{n-1} \neq j, X_n = j]$$

be the probability of a first visit to j at time n for a system that starts in i, and let

$$(8.15) \qquad f_{ij} = P_i\left(\bigcup_{n=1}^{\infty} [X_n = j]\right) = \sum_{n=1}^{\infty} f_{ij}^{(n)}$$

be the probability of an eventual visit. A state i is *persistent* if a system starting at i is certain sometime to return to i: $f_{ii} = 1$. The state is *transient* in the opposite case: $f_{ii} < 1$.

Suppose that $n_1, \ldots, n_k$ are integers satisfying $1 \leq n_1 < \cdots < n_k$ and consider the event that the system visits j at times $n_1 \ldots n_k$ but not in between; this event is determined by the conditions

$$\cdot \quad X_1 \neq j, \ldots, \qquad X_{n_1 - 1} \neq j, \qquad X_{n_1} = j,$$

$$X_{n_1 + 1} \neq j, \ldots, \qquad X_{n_2 - 1} \neq j, \qquad X_{n_2} = j,$$

$$\vdots$$

$$X_{n_{k-1} + 1} \neq j, \ldots, \qquad X_{n_k - 1} \neq j, \qquad X_{n_k} = j.$$

Repeated application of (8.13) shows that under P_i the probability of this event is $f_{ij}^{(n_1)} f_{jj}^{(n_2 - n_1)} \cdots f_{jj}^{(n_k - n_{k-1})}$. Add this over the k-tuples $n_1, \ldots, n_k$: the P_i-probability that $X_n = j$ for at least k different values of n is $f_{ij} f_{jj}^{k-1}$. Letting $k \to \infty$ therefore gives

$$(8.16) \qquad P_i[X_n = j \text{ i.o.}] = \begin{cases} 0 & \text{if } f_{jj} < 1, \\ f_{ij} & \text{if } f_{jj} = 1. \end{cases}$$

Recall that *i.o.* means *infinitely often*. Taking $i = j$ gives

$$(8.17) \qquad P_i[X_n = i \text{ i.o.}] = \begin{cases} 0 & \text{if } f_{ii} < 1, \\ 1 & \text{if } f_{ii} = 1. \end{cases}$$

Thus $P_i[X_n = i \text{ i.o.}]$ is either 0 or 1; compare the zero–one law (Theorem 4.5), but note that the events $[X_n = i]$ here are not in general independent.[†]

[†]See Problem 8.35

Theorem 8.2.

(i) *Transience of i is equivalent to* $P_i[X_n = i \text{ i.o.}] = 0$ *and to* $\sum_n p_{ii}^{(n)} < \infty$.

(ii) *Persistence of i is equivalent to* $P_i[X_n = i \text{ i.o.}] = 1$ *and to* $\sum_n p_{ii}^{(n)} = \infty$.

PROOF.   By the first Borel-Cantelli lemma, $\sum_n p_{ii}^{(n)} < \infty$ implies $P_i[X_n = i$ i.o.] = 0, which by (8.17) in turn implies $f_{ii} < 1$. The entire theorem will be proved if it is shown that $f_{ii} < 1$ implies $\sum_n p_{ii}^{(n)} < \infty$.

The proof uses a first-passage argument: By (8.13),

$$p_{ij}^{(n)} = P_i[X_n = j] = \sum_{s=0}^{n-1} P_i[X_1 \neq j, \ldots, X_{n-s-1} \neq j, X_{n-s} = j, X_n = j]$$

$$= \sum_{s=0}^{n-1} P_i[X_1 \neq j, \ldots, X_{n-s-1} \neq j, X_{n-s} = j] P_j[X_s = j]$$

$$= \sum_{s=0}^{n-1} f_{ij}^{(n-s)} p_{jj}^{(s)}.$$

Therefore,

$$\sum_{t=1}^{n} p_{ii}^{(t)} = \sum_{t=1}^{n} \sum_{s=0}^{t-1} f_{ii}^{(t-s)} p_{ii}^{(s)}$$

$$= \sum_{s=0}^{n-1} p_{ii}^{(s)} \sum_{t=s+1}^{n} f_{ii}^{(t-s)} \leq \sum_{s=0}^{n} p_{ii}^{(s)} f_{ii}.$$

Thus $(1 - f_{ii}) \sum_{t=1}^{n} p_{ii}^{(t)} \leq f_{ii}$, and if $f_{ii} < 1$, this puts a bound on the partial sums $\sum_{t=1}^{n} p_{ii}^{(t)}$.   ■

Example 8.6.   *Pólya's theorem.* For the symmetric k-dimensional random walk (Example 8.4), all states are persistent if $k = 1$ or $k = 2$, and all states are transient if $k \geq 3$. To prove this, note first that the probability $p_{ii}^{(n)}$ of return in n steps is the same for all states i; denote this probability by $a_n^{(k)}$ to indicate the dependence on the dimension k. Clearly, $a_{2n+1}^{(k)} = 0$. Suppose that $k = 1$. Since return in $2n$ steps means n steps east and n steps west,

$$a_{2n}^{(1)} = \binom{2n}{n} \frac{1}{2^{2n}}.$$

By Stirling's formula, $a_{2n}^{(1)} \sim (\pi n)^{-1/2}$. Therefore, $\sum_n a_n^{(1)} = \infty$, and all states are persistent by Theorem 8.2.

In the plane, a return to the starting point in $2n$ steps means equal numbers of steps east and west as well as equal numbers north and south:

$$a_{2n}^{(2)} = \sum_{u=0}^{n} \frac{(2n)!}{u!u!(n-u)!(n-u)!} \frac{1}{4^{2n}}$$

$$= \frac{1}{4^{2n}} \binom{2n}{n} \sum_{u=0}^{n} \binom{n}{u}\binom{n}{n-u}.$$

It can be seen on combinatorial grounds that the last sum is $\binom{2n}{n}$, and so $a_{2n}^{(2)} = (a_{2n}^{(1)})^2 \sim (\pi n)^{-1}$. Again, $\sum_n a_n^{(2)} = \infty$ and every state is persistent.

For three dimensions,

$$a_{2n}^{(3)} = \sum \frac{(2n)!}{u!u!v!v!(n-u-v)!(n-u-v)!} \frac{1}{6^{2n}},$$

the sum extending over nonnegative u and v satisfying $u+v \le n$. This reduces to

$$(8.18) \qquad a_{2n}^{(3)} = \sum_{i=0}^{n} \binom{2n}{2l}\left(\frac{1}{3}\right)^{2n-2l}\left(\frac{2}{3}\right)^{2l} a_{2n-2l}^{(1)} a_{2l}^{(2)},$$

as can be checked by substitution. (To see the probabilistic meaning of this formula, condition on there being $2n - 2l$ steps parallel to the vertical axis and $2l$ steps parallel to the horizontal plane.) It will be shown that $a_{2n}^{(3)} = O(n^{-3/2})$, which will imply that $\sum_n a_n^{(3)} < \infty$. The terms in (8.18) for $l = 0$ and $l = n$ are each $O(n^{-3/2})$ and hence can be omitted. Now $a_u^{(1)} \le Ku^{-1/2}$ and $a_u^{(2)} \le Ku^{-1}$, as already seen, and so the sum in question is at most

$$K^2 \sum_{l=1}^{n-1} \binom{2n}{2l}\left(\frac{1}{3}\right)^{2n-2l}\left(\frac{2}{3}\right)^{2l}(2n-2l)^{-1/2}(2l)^{-1}.$$

Since $(2n - 2l)^{-1/2} \le 2n^{1/2}(2n-2l)^{-1} \le 4n^{1/2}(2n-2l+1)^{-1}$ and $(2l)^{-1} \le 2(2l+1)^{-1}$, this is at most a constant times

$$n^{1/2}\frac{(2n)!}{(2n+2)!}\sum_{l=1}^{n-1}\binom{2n+2}{2l-1}\left(\frac{1}{3}\right)^{2n-2l+1}\left(\frac{2}{3}\right)^{2l+1} = O(n^{-3/2}).$$

Thus $\sum_n a_n^{(3)} < \infty$, and the states are transient. The same is true for $k = 4, 5, \ldots$, since an inductive extension of the argument shows that $a_n^{(k)} = O(n^{-k/2})$. ∎

It is possible for a system starting in i to reach j ($f_{ij} > 0$) if and only if $p_{ij}^{(n)} > 0$ for some n. If this is true for all i and j, the Markov chain is *irreducible*.

Theorem 8.3. *If the Markov chain is irreducible, then one of the following two alternatives holds.*

(i) *All states are transient*, $P_i(\bigcup_j [X_n = j \text{ i.o.}]) = 0$ *for all* i, *and* $\sum_n p_{ij}^{(n)} < \infty$ *for all* i *and* j.

(ii) *All states are persistent*, $P_i(\bigcap_j [X_n = j \text{ i.o.}]) = 1$ *for all* i, *and* $\sum_n p_{ij}^{(n)} = \infty$ *for all* i *and* j.

The irreducible chain itself can accordingly be called persistent or transient. In the persistent case the system visits every state infinitely often. In the transient case it visits each state only finitely often, hence visits each finite set only finitely often, and so may be said to go to infinity.

PROOF. For each i and j there exist r and s such that $p_{ij}^{(r)} > 0$ and $p_{ji}^{(s)} > 0$. Now

$$(8.19) \qquad\qquad p_{ii}^{(r+s+n)} \geq p_{ij}^{(r)} p_{jj}^{(n)} p_{ji}^{(s)},$$

and from $p_{ij}^{(r)} p_{ji}^{(s)} > 0$ it follows that $\sum_n p_{ii}^{(n)} < \infty$ implies $\sum_n p_{jj}^{(n)} < \infty$: if one state is transient, they all are. In this case (8.16) gives $P_i[X_n = j \text{ i.o.}] = 0$ for all i and j, so that $P_i(\bigcup_j [X_n = j \text{ i.o.}]) = 0$ for all i. Since $\sum_{n=1}^{\infty} p_{ij}^{(n)} = \sum_{n=1}^{\infty} \sum_{\nu=1}^{n} f_{ij}^{(\nu)} p_{jj}^{(n-\nu)} = \sum_{\nu=1}^{\infty} f_{ij}^{(\nu)} \sum_{m=0}^{\infty} p_{jj}^{(m)} \leq \sum_{m=0}^{\infty} p_{jj}^{(m)}$, it follows that if j is transient, then (Theorem 8.2) $\sum_n p_{ij}^{(n)}$ converges for every i.

The other possibility is that all states are persistent. In this case $P_j[X_n = j \text{ i.o.}] = 1$ by Theorem 8.2, and it follows by (8.13) that

$$p_{ji}^{(m)} = P_j([X_m = i] \cap [X_n = j \text{ i.o.}])$$

$$\leq \sum_{n > m} P_j[X_m = i, X_{m+1} \neq j, \ldots, X_{n-1} \neq j, X_n = j]$$

$$= \sum_{n > m} p_{ji}^{(m)} f_{ij}^{(n-m)} = p_{ji}^{(m)} f_{ij}.$$

There is an m for which $p_{ji}^{(m)} > 0$, and therefore $f_{ij} = 1$. By (8.16), $P_i[X_n = j \text{ i.o.}] = f_{ij} = 1$. If $\sum_n p_{ij}^{(n)}$ were to converge for some i and j, it would follow by the first Borel–Cantelli lemma that $P_i[X_n = j \text{ i.o.}] = 0$. ∎

Example 8.7. Since $\sum_j p_{ij}^{(n)} = 1$, the first alternative in Theorem 8.3 is impossible if S is finite: a finite, irreducible Markov chain is persistent. ∎

Example 8.8. The chain in Pólya's theorem is certainly irreducible. If the dimension is 1 or 2, there is probability 1 that a particle in symmetric random walk visits every state infinitely often. If the dimension is 3 or more, the particle goes to infinity. ∎

Example 8.9. Consider the unrestricted random walk on the line (Example 8.3). According to the ruin calculation (7.8), $f_{01} = p/q$ for $p < q$. Since the chain is irreducible, all states are transient. By symmetry, of course, the chain is also transient if $p > q$, although in this case (7.8) gives $f_{01} = 1$. Thus $f_{ij} = 1$ $(i \neq j)$ is possible in the transient case.[†]

If $p = q = \frac{1}{2}$, the chain is persistent by Pólya's theorem. If n and $j - i$ have the same parity,

$$p_{ij}^{(n)} = \left(\begin{array}{c} n \\ \dfrac{n+j-i}{2} \end{array} \right) \frac{1}{2^n}, \qquad |j - i| \leq n.$$

This is maximal if $j = i$ or $j = i \pm 1$, and by Stirling's formula the maximal value is of order $n^{-1/2}$. Therefore, $\lim_n p_{ij}^{(n)} = 0$, which always holds in the transient case but is thus possible in the persistent case as well (see Theorem 8.8).   ∎

Another Criterion for Persistence

Let $Q = [q_{ij}]$ be a matrix with rows and columns indexed by the elements of a finite or countable set U. Suppose it is *substochastic* in the sense that $q_{ij} \geq 0$ and $\sum_j q_{ij} \leq 1$. Let $Q^n = [q_{ij}^{(n)}]$ be the nth power, so that

$$(8.20) \qquad q_{ij}^{(n+1)} = \sum_\nu q_{i\nu} q_{\nu j}^{(n)}, \qquad q_{ij}^{(0)} = \delta_{ij}.$$

Consider the row sums

$$(8.21) \qquad \sigma_i^{(n)} = \sum_j q_{ij}^{(n)}.$$

From (8.20) follows

$$(8.22) \qquad \sigma_i^{(n+1)} = \sum_j q_{ij} \sigma_j^{(n)}.$$

Since Q is substochastic $\sigma_i^{(1)} \leq 1$, and hence $\sigma_i^{(n+1)} = \sum_j \sum_\nu q_{i\nu}^{(n)} q_{\nu j} = \sum_\nu q_{i\nu}^{(n)} \sigma_\nu^{(1)} \leq \sigma_i^{(n)}$. Therefore, the monotone limits

$$(8.23) \qquad \sigma_i = \lim_n \sum_j q_{ij}^{(n)}$$

[†] But for each j there must be some $i \neq j$ for which $f_{ij} < 1$; see Problem 8.7.

exist. By (8.22) and the Weierstrass M-test [A28], $\sigma_i = \sum_j q_{ij}\sigma_j$. Thus the σ_i solve the system

(8.24)
$$\begin{cases} x_i = \sum_{j \in U} q_{ij} x_j, & i \in U, \\ 0 \le x_i \le 1, & i \in U. \end{cases}$$

For an arbitrary solution, $x_i = \sum_j q_{ij} x_j \le \sum_j q_{ij} = \sigma_i^{(1)}$, and $x_i \le \sigma_i^{(n)}$ for all i implies $x_i \le \sum_j q_{ij}\sigma_j^{(n)} = \sigma_i^{(n+1)}$ by (8.22). Thus $x_i \le \sigma_i^{(n)}$ for all n by induction, and so $x_i \le \sigma_i$. Thus the σ_i give the *maximal* solution to (8.24):

Lemma 1. *For a substochastic matrix Q the limits (8.23) are the maximal solution of (8.24).*

Now suppose that U is a subset of the state space S. The p_{ij} for i and j in U give a substochastic matrix Q. The row sums (8.21) are $\sigma_i^{(n)} = \sum p_{ij_1} p_{j_1 j_2}$ $\cdots p_{j_{n-1} j_n}$, where the $j_1, \ldots, j_n$ range over U, and so $\sigma_i^{(n)} = P_i[X_t \in U, t \le n]$. Let $n \to \infty$:

(8.25)
$$\sigma_i = P_i[X_t \in U, t = 1, 2 \ldots], \qquad i \in U.$$

In this case, σ_i is thus the probability that the system remains forever in U, given that it starts at i. The following theorem is now an immediate consequence of Lemma 1.

Theorem 8.4. *For $U \subset S$ the probabilities (8.25) are the maximal solution of the system*

(8.26)
$$\begin{cases} x_i = \sum_{j \in U} p_{ij} x_j, & i \in U, \\ 0 \le x_i \le 1, & i \in U. \end{cases}$$

The constraint $x_i \ge 0$ in (8.26) is in a sense redundant: Since $x_i \equiv 0$ is a solution, the maximal solution is automatically nonnegative (and similarly for (8.24)). And the maximal solution is $x_i \equiv 1$ if and only if $\sum_{j \in U} p_{ij} = 1$ for all i in U, which makes probabilistic sense.

Example 8.10. For the random walk on the line consider the set $U = \{0, 1, 2, \ldots\}$. The System (8.26) is

$$x_i = p x_{i+1} + q x_{i-1}, \qquad i \ge 1,$$
$$x_0 = p x_1.$$

It follows [A19] that $x_n = A + An$ if $p = q$ and $x_n = A - A(q/p)^{n+1}$ if $p \ne q$. The only bounded solution is $x_n \equiv 0$ if $q \ge p$, and in this case there is

probability 0 of staying forever among the nonnegative integers. If $q < p$, $A = 1$ gives the maximal solution $x_n = 1 - (q/p)^{n+1}$ (and $0 \leq A < 1$ gives exactly the solutions that are not maximal). Compare (7.8) and Example 8.9.

■

Now consider the system (8.26) with $U = S - \{i_0\}$ for an arbitrary single state i_0:

$$(8.27) \qquad \begin{cases} x_i = \sum_{j \neq i_0} p_{ij} x_j, & i \neq i_0, \\ 0 \leq x_i \leq 1, & i \neq i_0. \end{cases}$$

There is always the trivial solution—the one for which $x_i \equiv 0$.

Theorem 8.5. *An irreducible chain is transient if and only if (8.27) has a nontrivial solution.*

PROOF.   The probabilities

$$(8.28) \qquad 1 - f_{ii_0} = P_i[X_n \neq i_0, n \geq 1], \qquad i \neq i_0,$$

are by Theorem 8.4 the maximal solution of (8.27). Therefore (8.27) has a nontrivial solution if and only if $f_{ii_0} < 1$ for some $i \neq i_0$. If the chain is persistent, this is impossible by Theorem 8.3(ii).

Suppose the chain is transient. Since

$$f_{i_0 i_0} = P_{i_0}[X_1 = i_0] + \sum_{n=2}^{\infty} \sum_{i \neq i_0} P_{i_0}[X_1 = i, X_2 \neq i_0, \ldots, X_{n-1} \neq i_0, X_n = i_0]$$

$$= p_{i_0 i_0} + \sum_{i \neq i_0} p_{i_0 i} f_{i i_0},$$

and since $f_{i_0 i_0} < 1$, it follows that $f_{i i_0} < 1$ for some $i \neq i_0$.   ■

Since the equations in (8.27) are homogeneous, the issue is whether they have a solution that is nonnegative, nontrivial, and *bounded*. If they do, $0 \leq x_i \leq 1$ can be arranged by rescaling.[†]

[†]See Problem 8.9.

Example 8.11. In the simplest of *queueing models* the state space is $\{0,1,2,\ldots\}$ and the transition matrix has the form

$$\begin{bmatrix}
t_0 & t_1 & t_2 & 0 & 0 & 0 & \cdots \\
t_0 & t_1 & t_2 & 0 & 0 & 0 & \cdots \\
0 & t_0 & t_1 & t_2 & 0 & 0 & \cdots \\
0 & 0 & t_0 & t_1 & t_2 & 0 & \cdots \\
0 & 0 & 0 & t_0 & t_1 & t_2 & \cdots \\
& & & & & & \\
\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
& & & & & &
\end{bmatrix}$$

If there are i customers in the queue and $i \geq 1$, the customer at the head of the queue is served and leaves, and then 0, 1, or 2 new customers arrive (probabilities t_0, t_1, t_2), which leaves a queue of length $i - 1$, i, or $i + 1$. If $i = 0$, no one is served, and the new customers bring the queue length to 0, 1, or 2. Assume that t_0 and t_2 are positive, so that the chain is irreducible.

For $i_0 = 0$ the system (8.27) is

$$(8.29) \qquad \begin{aligned}
x_1 &= t_1 x_1 + t_2 x_2, \\
x_k &= t_0 x_{k-1} + t_1 x_k + t_2 x_{k+1}, \qquad k \geq 2.
\end{aligned}$$

Since t_0, t_1, t_2 have the form $q(1-t), t, p(1-t)$ for appropriate p, q, t, the second line of (8.29) has the form $x_k = p x_{k+1} + q x_{k-1}$, $k \geq 2$. Now the solution [A19] is $A + B(q/p)^k = A + B(t_0/t_2)^k$ if $t_0 \neq t_2$ $(p \neq q)$ and $A + Bk$ if $t_0 = t_2$ $(p = q)$, and A can be expressed in terms of B because of the first equation in (8.29). The result is

$$x_k = \begin{cases}
B\big((t_0/t_2)^k - 1\big) & \text{if } t_0 \neq t_2, \\
Bk & \text{if } t_0 = t_2.
\end{cases}$$

There is a nontrivial solution if $t_0 < t_2$ but not if $t_0 \geq t_2$.

If $t_0 < t_2$, the chain is thus transient, and the queue size goes to infinity with proability 1. If $t_0 \geq t_2$, the chain is persistent. For a nonempty queue the expected increase in queue length in one step is $t_2 - t_0$, and the queue goes out of control if and only if this is positive. ∎

Stationary Distributions

Suppose that the chain has initial probabilities π_i satisfying

$$(8.30) \qquad \sum_{i \in S} \pi_i p_{ij} = \pi_j, \qquad j \in S.$$

It then follows by induction that

$$(8.31) \qquad \sum_{i \in S} \pi_i p_{ij}^{(n)} = \pi_j, \qquad j \in S, \quad n = 0, 1, 2, \dots .$$

If π_i is the probability that $X_0 = i$, then the left side of (8.31) is the probability that $X_n = j$, and thus (8.30) implies that the probability of $[X_n = j]$ is the same for all n. A set of probabilities satisfying (8.30) is for this reason called a *stationary distribution*. The existence of such a distribution implies that the chain is very stable.

To discuss this requires the notion of periodicity. The state j has *period t* if $p_{jj}^{(n)} > 0$ implies that t divides n and if t is the largest integer with this property. In other words, the period of j is the greatest common divisor of the set of integers

$$(8.32) \qquad \left[n: n \geq 1, \, p_{jj}^{(n)} > 0 \right].$$

If the chain is irreducible, then for each pair i and j there exist r and s for which $p_{ij}^{(r)}$ and $p_{ji}^{(s)}$ are positive, and of course

$$(8.33) \qquad p_{ii}^{(r+s+n)} \geq p_{ij}^{(r)} p_{jj}^{(n)} p_{ji}^{(s)}.$$

Let t_i and t_j be the periods of i and j. Taking $n = 0$ in this inequality shows that t_i divides $r + s$; and now it follows by the inequality that $p_{jj}^{(n)} > 0$ implies that t_i divides $r + s + n$ and hence divides n. Thus t_i divides each integer in the set (8.32), and so $t_i \leq t_j$. Since i and j can be interchanged in this argument, i and j have the same period. One can thus speak of the period of the chain itself in the irreducible case. The random walk on the line has period 2, for example. If the period is 1, the chain is *aperiodic*

Lemma 2. *In an irreducible, aperiodic chain, for each i and j, $p_{ij}^{(n)} > 0$ for all n exceeding some $n_0(i, j)$.*

PROOF. Since $p_{jj}^{(m+n)} \geq p_{jj}^{(m)} p_{jj}^{(n)}$, if M is the set (8.32) then $m \in M$ and $n \in M$ together imply $m + n \in M$. But it is a fact of number theory [A21] that if a set of positive integers is closed under addition and has greatest common divisor 1, then it contains all integers exceeding some n_1. Given i and j, choose r so that $p_{ij}^{(r)} > 0$. If $n > n_0 = n_1 + r$, then $p_{ij}^{(n)} \geq p_{ij}^{(r)} p_{jj}^{(n-r)} > 0$.   ∎

Theorem 8.6. *Suppose of an irreducible, aperiodic chain that there exists a stationary distribution—a solution of (8.30) satisfying $\pi_j \geq 0$ and $\sum_j \pi_j = 1$. Then the chain is persistent,*

$$(8.34) \qquad \lim_n p_{ij}^{(n)} = \pi_j$$

for all i and j, the π_j are all positive, and the stationary distribution is unique.

The main point of the conclusion is that the effect of the initial state wears off. Whatever the actual initial distribution $\{\alpha_i\}$ of the chain may be, if (8.34) holds, then it follows by the M-test that the probability $\sum_i \alpha_i p_{ij}^{(n)}$ of $[X_n = j]$ converges to π_j.

PROOF. If the chain is transient, then $p_{ij}^{(n)} \to 0$ for all i and j by Theorem 8.3, and it follows by (8.31) and the M-test that π_j is identically 0, which contradicts $\sum_i \pi_i = 1$. The existence of a stationary distribution therefore implies that the chain is persistent.

Consider now a Markov chain with state space $S \times S$ and transition probabilities $p(ij, kl) = p_{ik} p_{jl}$ (it is easy to verify that these form a stochastic matrix). Call this the *coupled* chain; it describes the joint behavior of a pair of independent systems, each evolving according to the laws of the original Markov chain. By Theorem 8.1 there exists a Markov chain (X_n, Y_n), $n = 0, 1, \ldots$, having positive initial probabilities and transition probabilities

$$P\big[(X_{n+1}, Y_{n+1}) = (k, l) \,|\, (X_n, Y_n) = (i, j)\big] = p(ij, kl).$$

For n exceeding some n_0 depending on $i. j, k, l$, the probability $p^{(n)}(ij, kl) = p_{ik}^{(n)} p_{jl}^{(n)}$ is positive by Lemma 2. Therefore, the coupled chain is *irreducible*. (This proof that the coupled chain is irreducible requires only the assumptions that the original chain is irreducible and aperiodic, a fact needed again in the proof of Theorem 8.7.)

It is easy to check that $\pi(ij) = \pi_i \pi_j$ forms a set of stationary initial probabilities for the coupled chain, which, like the original one, must therefore be *persistent*. It follows that, for an arbitrary initial state (i, j) for the chain $\{(X_n, Y_n)\}$ and an arbitrary i_0 in S, one has $P_{ij}[(X_n, Y_n) = (i_0, i_0)$ i.o.] = 1. If τ is the smallest integer such that $X_\tau = Y_\tau = i_0$, then τ is finite with probability 1 under P_{ij}. The idea of the proof is now this: X_n starts in i and Y_n starts in j; once $X_n = Y_n = i_0$ occurs, X_n and Y_n follow identical probability laws, and hence the initial states i and j will lose their influence.

By (8.13) applied to the coupled chain, if $m \le n$, then

$$P_{ij}\big[(X_n, Y_n) = (k, l), \tau = m\big]$$

$$= P_{ij}\big[(X_t, Y_t) \ne (i_0, i_0), t < m, (X_m, Y_m) = (i_0, i_0)\big]$$

$$\times P_{i_0 i_0}\big[(X_{n-m}, Y_{n-m}) = (k, l)\big]$$

$$= P_{ij}[\tau = m] p_{i_0 k}^{(n-m)} p_{i_0 l}^{(n-m)}.$$

Adding out l gives $P_{ij}[X_n = k, \tau = m] = P_{ij}[\tau = m] p_{i_0 k}^{(n-m)}$, and adding out k gives $P_{ij}[Y_n = l, \tau = m] = P_{ij}[\tau = m] p_{i_0 l}^{(n-m)}$. Take $k = l$, equate probabilities, and add over $m = 1, \ldots, n$:

$$P_{ij}[X_n = k, \tau \le n] = P_{ij}[Y_n = k, \tau \le n].$$

From this follows

$$P_{ij}[X_n = k] \leq P_{ij}[X_n = k, \tau \leq n] + P_{ij}[\tau > n]$$

$$= P_{ij}[Y_n = k, \tau \leq n] + P_{ij}[\tau > n]$$

$$\leq P_{ij}[Y_n = k] + P_{ij}[\tau > n].$$

This and the same inequality with X and Y interchanged give

$$\left| p_{ik}^{(n)} - p_{jk}^{(n)} \right| = \left| P_{ij}[X_n = k] - P_{ij}[Y_n = k] \right| \leq P_{ij}[\tau > n].$$

Since τ is finite with probability 1,

$$(8.35) \qquad \qquad \lim_n \left| p_{ik}^{(n)} - p_{jk}^{(n)} \right| = 0.$$

(This proof of (8.35) goes through as long as the coupled chain is irreducible and persistent— no assumptions on the original chain are needed. This fact is used in the proof of the next theorem.)

By (8.31), $\pi_k - p_{jk}^{(n)} = \sum_i \pi_i (p_{ik}^{(n)} - p_{jk}^{(n)})$, and this goes to 0 by the M-test if (8.35) holds. Thus $\lim_n p_{jk}^{(n)} = \pi_k$. As this holds for each stationary distribution, there can be only one of them.

It remains to show that the π_j are all strictly positive. Choose r and s so that $p_{ij}^{(r)}$ and $p_{ji}^{(s)}$ are positive. Letting $n \to \infty$ in (8.33) shows that π_i is positive if π_j is; since some π_j is positive (they add to 1), all the π_i must be positive. ∎

Example 8.12. For the queueing model in Example 8.11 the equations (8.30) are

$$\pi_0 = \pi_0 t_0 + \pi_1 t_0,$$

$$\pi_1 = \pi_0 t_1 + \pi_1 t_1 + \pi_2 t_0,$$

$$\pi_2 = \pi_0 t_2 + \pi_1 t_2 + \pi_2 t_1 + \pi_3 t_0,$$

$$\pi_k = \pi_{k-1} t_2 + \pi_k t_1 + \pi_{k+1} t_0, \qquad k \geq 3.$$

Again write t_0, t_1, t_2, as $q(1-t), t, p(1-t)$. Since the last equation here is $\pi_k = q\pi_{k+1} + p\pi_{k-1}$, the solution is

$$\pi_k = \begin{cases} A + B(p/q)^k = A + B(t_2/t_0)^k & \text{if } t_0 \neq t_2, \\ A + Bk & \text{if } t_0 = t_2 \end{cases}$$

for $k \geq 2$. If $t_0 < t_2$ and $\sum \pi_k$ converges, then $\pi_k \equiv 0$, and hence there is no stationary distribution; but this is not new, because it was shown in Example 8.11 that the chain is transient in this case. If $t_0 = t_2$, there is again no

stationary distribution, and this is new because the chain was in Example 8.11 shown to be persistent in this case.

If $t_0 > t_2$, then $\Sigma \pi_k$ converges, provided $A = 0$. Solving for π_0 and π_1 in the first two equations of the system above gives $\pi_0 = Bt_2$ and $\pi_1 = Bt_2(1 - t_0)/t_0$. From $\Sigma_k \pi_k = 1$ it now follows that $B = (t_0 - t_2)/t_2$, and the π_k can be written down explicitly. Since $\pi_k = B(t_2/t_0)^k$ for $k \geq 2$, there is small chance of a large queue length.   ∎

If $t_0 = t_2$ in this queueing model, the chain is persistent (Example 8.11) but has no stationary distribution (Example 8.12). The next theorem describes the asymptotic behavior of the $p_{ij}^{(n)}$ in this case.

Theorem 8.7. *If an irreducible, aperiodic chain has no stationary distribution, then*

$$(8.36) \qquad\qquad\qquad \lim_n p_{ij}^{(n)} = 0$$

for all i and j.

If the chain is transient, (8.36) follows from Theorem 8.3. What is interesting here is the persistent case.

PROOF. By the argument in the proof of Theorem 8.6, the coupled chain is irreducible. If it is transient, then $\Sigma_n (p_{ij}^{(n)})^2$ converges by Theorem 8.2, and the conclusion follows.

Suppose, on the other hand, that the coupled chain is (irreducible and) persistent. Then the stopping-time argument leading to (8.35) goes through as before. If the $p_{ij}^{(n)}$ do not all go to 0, then there is an increasing sequence $\{n_u\}$ of integers along which some $p_{ij}^{(n)}$ is bounded away from 0. By the diagonal method [A14], it is possible by passing to a subsequence of $\{n_u\}$ to ensure that each $p_{ij}^{(n_u)}$ converges to a limit, which by (8.35) must be independent of i. Therefore, there is a sequence $\{n_u\}$ such that $\lim_u p_{ij}^{(n_u)} = t_j$ exists for all i and j, where t_j is nonnegative for all j and positive for some j. If M is a finite set of states, then $\Sigma_{j \in M} t_j = \lim_u \Sigma_{j \in M} p_{ij}^{(n_u)} \leq 1$, and hence $0 < t = \Sigma_j t_j \leq 1$. Now $\Sigma_{k \in M} p_{ik}^{(n_u)} p_{kj} \leq p_{ij}^{(n_u + 1)} = \Sigma_k p_{ik} p_{kj}^{(n_u)}$; it is possible to pass to the limit ($u \to \infty$) inside the first sum (if M is finite) and inside the second sum (by the M-test), and hence $\Sigma_{k \in M} t_k p_{kj} \leq \Sigma_k p_{ik} t_j = t_j$. Therefore, $\Sigma_k t_k p_{kj} \leq t_j$; if one of these inequalities were strict, it would follow that $\Sigma_k t_k = \Sigma_j \Sigma_k t_k p_{kj} < \Sigma_j t_j$, which is impossible. Therefore $\Sigma_k t_k p_{kj} = t_j$ for all j, and the ratios $\pi_j = t_j/t$ give a stationary distribution, contrary to the hypothesis.   ∎

The limits in (8.34) and (8.36) can be described in terms of mean return times. Let

$$(8.37) \qquad\qquad\qquad \mu_j = \sum_{n=1}^{\infty} n f_{jj}^{(n)};$$

if the series diverges, write $\mu_j = \infty$. In the persistent case, this sum is to be thought of as the average number of steps to first return to j, given that $X_0 = j$.[†]

Lemma 3.   *Suppose that j is persistent and that $\lim_n p_{jj}^{(n)} = u$. Then $u > 0$ if and only if $\mu_j < \infty$, in which case $u = 1/\mu_j$.*

Under the convention that $0 = 1/\infty$, the case $u = 0$ and $\mu_j = \infty$ is consistent with the equation $u = 1/\mu_j$.

PROOF.   For $k \geq 0$ let $\rho_k = \sum_{n > k} f_{jj}^{(n)}$; the notation does not show the dependence on j, which is fixed. Consider the double series

$$f_{jj}^{(1)} + f_{jj}^{(2)} + f_{jj}^{(3)} + \cdots$$
$$+ f_{jj}^{(2)} + f_{jj}^{(3)} + \cdots$$
$$+ f_{jj}^{(3)} + \cdots$$
$$+ \cdots$$

The kth row sums to ρ_k $(k \geq 0)$ and the nth column sums to $n f_{jj}^{(n)}$ $(n \geq 1)$, and so [A27] the series in (8.37) converges if and only if $\sum_k \rho_k$ does, in which case

$$(8.38) \qquad\qquad \mu_j = \sum_{k=0}^{\infty} \rho_k.$$

Since j is persistent, the P_j-probability that the system does not hit j up to time n is the probability that it hits j after time n, and this is ρ_n. Therefore,

$$1 - p_{jj}^{(n)} = P_j[X_n \neq j]$$

$$= P_j[X_1 \neq j, \ldots, X_n \neq j] + \sum_{k=1}^{n-1} P_j[X_k = j, X_{k+1} \neq j, \ldots, X_n \neq j]$$

$$= \rho_n + \sum_{k=1}^{n-1} p_{jj}^{(k)} \rho_{n-k},$$

and since $\rho_0 = 1$,

$$1 = \rho_0 p_{jj}^{(n)} + \rho_1 p_{jj}^{(n-1)} + \cdots + \rho_{n-1} p_{jj}^{(1)} + \rho_n p_{jj}^{(0)}.$$

Keep only the first $k + 1$ terms on the right here, and let $n \to \infty$; the result is $1 \geq (\rho_0 + \cdots + \rho_k)u$. Therefore $u > 0$ implies that $\sum_k \rho_k$ converges, so that $\mu_j < \infty$.

[†]Since in general there is no upper bound to the number of steps to first return, it is not a simple random variable. It does come under the general theory in Chapter 4, and its expected value is indeed μ_j (and (8 38) is just (5.29)), but for the present the interpretation of μ_j as an average is informal See Problem 23.11

Write $x_{nk} = \rho_k p_{jj}^{(n-k)}$ for $0 \leq k \leq n$ and $x_{nk} = 0$ for $n < k$. Then $0 \leq x_{nk} \leq \rho_k$ and $\lim_n x_{nk} = \rho_k u$. If $\mu_j < \infty$, then $\Sigma_k \rho_k$ converges and it follows by the M-test that $1 = \Sigma_{k=0}^{\infty} x_{nk} \to \Sigma_{k=0}^{\infty} \rho_k u$. By (8.38), $1 = \mu_j u$, so that $u > 0$ and $u = 1/\mu_j$. ∎

The law of large numbers bears on the relation $u = 1/\mu_j$ in the persistent case. Let V_n be the number of visits to state j up to time n. If the time from one visit to the next is about μ_j, then V_n should be about n/μ_j: $V_n/n \approx 1/\mu_j$. But (if $X_0 = j$) V_n/n has expected value $n^{-1}\Sigma_{k=1}^{n} p_{jj}^{(k)}$, which goes to u under the hypothesis of Lemma 3 [A30].

Consider an irreducible, aperiodic, persistent chain. There are two possibilities. If there is a stationary distribution, then the limits (8.34) are positive, and the chain is called *positive persistent*. It then follows by Lemma 3 that $\mu_j < \infty$ and $\pi_j = 1/\mu_j$ for all j. In this case, it is not actually necessary to assume persistence, since this follows from the existence of a stationary distribution. On the other hand, if the chain has no stationary distribution, then the limits (8.36) are all 0, and the chain is called *null persistent* It then follows by Lemma 3 that $\mu_j = \infty$ for all j. This, taken together with Theorem 8.3, provides a complete classification:

Theorem 8.8. *For an irreducible, aperiodic chain there are three possibilities*:

(i) *The chain is transient*; *then for all i and j,* $\lim_n p_{ij}^{(n)} = 0$ *and in fact* $\Sigma_n p_{ij}^{(n)} < \infty$.

(ii) *The chain is persistent but there exists no stationary distribution* (*the null persistent case*); *then for all i and j,* $p_{ij}^{(n)}$ *goes to 0 but so slowly that* $\Sigma_n p_{ij}^{(n)} = \infty$, *and* $\mu_j = \infty$.

(iii) *There exist stationary probabilities π_j and* (*hence*) *the chain is persistent* (*the positive persistent case*); *then for all i and j,* $\lim_n p_{ij}^{(n)} = \pi_j > 0$ *and* $\mu_j = 1/\pi_j < \infty$.

Since the asymptotic properties of the $p_{ij}^{(n)}$ are distinct in the three cases, these asymptotic properties in fact characterize the three cases.

Example 8.13. Suppose that the states are $0, 1, 2, \ldots$ and the transition matrix is

$$\begin{bmatrix} q_0 & p_0 & 0 & 0 & \cdots \\ q_1 & 0 & p_1 & 0 & \cdots \\ q_2 & 0 & 0 & p_2 & \cdots \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix}$$

where p_i and q_i are positive. The state i represents the length of a success

run, the conditional chance of a further success being p_i. Clearly the chain is irreducible and aperiodic.

A solution of the system (8.27) for testing for transience (with $i_0 = 0$) must have the form $x_k = x_1/p_1 \cdots p_{k-1}$. Hence there is a bounded, nontrivial solution, and the chain is transient, if and only if the limit α of $p_0 \cdots p_n$ is positive. But the chance of no return to 0 (for initial state 0) in n steps is clearly $p_0 \cdots p_{n-1}$; hence $f_{00} = 1 - \alpha$, which checks: the chain is persistent if and only if $\alpha = 0$.

Every solution of the steady-state equations (8.30) has the form $\pi_k = \pi_0 p_0 \cdots p_{k-1}$. Hence there is a stationary distribution if and only if $\Sigma_k p_0 \cdots p_k$ converges; this is the positive persistent case. The null persistent case is that in which $p_0 \cdots p_k \to 0$ but $\Sigma_k p_0 \cdots p_k$ diverges (which happens, for example, if $q_k = 1/k$ for $k > 1$).

Since the chance of no return to 0 in n steps is $p_0 \cdots p_{n-1}$, in the persistent case (8.38) gives $\mu_0 = \Sigma_{k=0}^{\infty} p_0 \cdots p_{k-1}$. In the null persistent case this checks with $\mu_0 = \infty$; in the positive persistent case it gives $\mu_0 = \Sigma_{k=0}^{\infty} \pi_k/\pi_0 = 1/\pi_0$, which again is consistent. ∎

Example 8.14. Since $\Sigma_j p_{ij}^{(n)} = 1$, possibilities (i) and (ii) in Theorem 8.8 are impossible in the finite case: A finite, irreducible, aperiodic Markov chain has a stationary distribution. ∎

Exponential Convergence*

In the finite case, $p_{ij}^{(n)}$ converges to π_j at an exponential rate:

Theorem 8.9.   *If the state space is finite and the chain is irreducible and aperiodic, then there is a stationary distribution $\{\pi_i\}$, and*

$$\left| p_{ij}^{(n)} - \pi_j \right| \leq A\rho^n,$$

where $A \geq 0$ and $0 \leq \rho < 1$.

PROOF.[†]   Let $m_j^{(n)} = \min_i p_{ij}^{(n)}$ and $M_j^{(n)} = \max_i p_{ij}^{(n)}$. By (8.10),

$$m_j^{(n+1)} = \min_i \sum_\nu p_{i\nu} p_{\nu j}^{(n)} \geq \min_i \sum_\nu p_{i\nu} m_j^{(n)} = m_j^{(n)},$$

$$M_j^{(n+1)} = \max_i \sum_\nu p_{i\nu} p_{\nu j}^{(n)} \leq \max_i \sum_\nu p_{i\nu} M_j^{(n)} = M_j^{(n)}.$$

---

*This topic may be omitted.
[†]For other proofs, see Problems 8.18 and 8 27.

Since obviously $m_j^{(n)} \le M_j^{(n)}$,

(8.39) $$0 \le m_j^{(1)} \le m_j^{(2)} \le \cdots \le M_j^{(2)} \le M_j^{(1)} \le 1.$$

Suppose temporarily that all the p_{ij} are positive. Let s be the number of states and let $\delta = \min_{ij} p_{ij}$. From $\Sigma_j p_{ij} \ge s\delta$ follows $0 < \delta \le s^{-1}$. Fix states u and v for the moment; let Σ' denote the summation over j in S satisfying $p_{uj} \ge p_{vj}$ and let Σ'' denote summation over j satisfying $p_{uj} < p_{vj}$. Then

(8.40) $$\Sigma'(p_{uj} - p_{vj}) + \Sigma''(p_{uj} - p_{vj}) = 1 - 1 = 0.$$

Since $\Sigma' p_{vj} + \Sigma'' p_{uj} \ge s\delta$.

(8.41) $$\Sigma'(p_{uj} - p_{vj}) = 1 - \Sigma'' p_{uj} - \Sigma' p_{vj} \le 1 - s\delta.$$

Apply (8.40) and then (8.41):

$$p_{uk}^{(n+1)} - p_{vk}^{(n+1)} = \sum_j (p_{uj} - p_{vj}) p_{jk}^{(n)}$$

$$\le \Sigma'(p_{uj} - p_{vj}) M_k^{(n)} + \Sigma''(p_{uj} - p_{vj}) m_k^{(n)}$$

$$= \Sigma'(p_{uj} - p_{vj})(M_k^{(n)} - m_k^{(n)})$$

$$\le (1 - s\delta)(M_k^{(n)} - m_k^{(n)}).$$

Since u and v are arbitrary,

$$M_k^{(n+1)} - m_k^{(n+1)} \le (1 - s\delta)(M_k^{(n)} - m_k^{(n)}).$$

Therefore, $M_k^{(n)} - m_k^{(n)} \le (1 - s\delta)^n$. It follows by (8.39) that $m_j^{(n)}$ and $M_j^{(n)}$ have a common limit π_j and that

(8.42) $$\left| p_{ij}^{(n)} - \pi_j \right| \le (1 - s\delta)^n.$$

Take $A = 1$ and $\rho = 1 - s\delta$. Passing to the limit in $\Sigma_i p_{vi}^{(n)} p_{ij} = p_{vj}^{(n+1)}$ shows that the π_i are stationary probabilities. (Note that the proof thus far makes almost no use of the preceding theory.)

If the p_{ij} are not all positive, apply Lemma 2: Since there are only finitely many states, there exists an m such that $p_{ij}^{(m)} > 0$ for all i and j. By the case just treated, $M_j^{(ml)} - m_j^{(ml)} \le \rho^l$. Take $A = \rho^{-1}$ and then replace ρ by $\rho^{1/m}$.

■

Example 8.15.   Suppose that

$$P = \begin{bmatrix} p_0 & p_1 & \cdots & p_{s-1} \\ p_{s-1} & p_0 & \cdots & p_{s-2} \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ p_1 & p_2 & \cdots & p_0 \end{bmatrix}.$$

The rows of P are the cyclic permutations of the first row: $p_{ij} = p_{j-i}$, $j - i$ reduced modulo s. Since the columns of P add to 1 as well as the rows, the steady-state equations (8.30) have the solution $\pi_i \equiv s^{-1}$. If the p_i are all positive, the theorem implies that $p_{ij}^{(n)}$ converges to s^{-1} at an exponential rate. If $X_0, Y_1, Y_2, \ldots$ are independent random variables with range $\{0, 1, \ldots, s - 1\}$, if each Y_n has distribution $\{p_0, \ldots, p_{s-1}\}$, and if $X_n = X_0 + Y_1 + \cdots + Y_n$, where the sum is reduced modulo s, then $P[X_n = j] \to s^{-1}$. The X_n describe a random walk on a circle of points, and whatever the initial distribution, the positions become equally likely in the limit.   ∎

Optimal Stopping*

Assume throughout the rest of the section that S *is finite*. Consider a function τ on Ω for which $\tau(\omega)$ is a nonnegative integer for each ω. Let $\mathscr{F}_n = \sigma(X_0, X_1, \ldots, X_n)$; τ is a *stopping time* or a *Markov time* if

(8.43) $\left[\omega : \tau(\omega) = n \right] \in \mathscr{F}_n$

for $n = 0, 1, \ldots$. This is analogous to the condition (7.18) on the gambler's stopping time. It will be necessary to allow $\tau(\omega)$ to assume the special value ∞, but only on a set of probability 0. This has no effect on the requirement (8.43), which concerns finite n only.

If f is a real function on the state space, then $f(X_0), f(X_1), \ldots$ are simple random variables. Imagine an observer who follows the successive states $X_0, X_1, \ldots$ of the system. He stops at time τ, when the state is X_τ (or $X_{\tau(\omega)}(\omega)$), and receives an reward or payoff $f(X_\tau)$. The condition (8.43) prevents prevision on the part of the observer. This is a kind of game, the stopping time is a strategy, and the problem is to find a strategy that maximizes the expected payoff $E[f(X_\tau)]$. The problem in Example 8.5 had this form; there $S = \{1, 2, \ldots, r + 1\}$, and the payoff function is $f(i) = i/r$ for $i \le r$ (set $f(r + 1) = 0$).

If $P(A) > 0$ and $Y = \sum_j y_j I_{B_j}$ is a simple random variable, the B_j forming a finite decomposition of Ω into $\mathscr{F}$-sets, the conditional expected value of Y

---

*This topic may be omitted.

given A is defined by

$$E[Y|A] = \sum_j y_i P(B_j|A).$$

Denote by E_i conditional expected values for the case $A = [X_0 = i]$:

$$E_i[Y] = E[Y|X_0 = i] = \sum_j y_j P_i(B_j).$$

The stopping-time problem is to choose τ so as to maximize simultaneously $E_i[f(X_\tau)]$ for all initial states i. If x lies in the range of f, which is finite, and if τ is everywhere finite, then $[\omega: f(X_{\tau(\omega)}(\omega)) = x] = \bigcup_{n=0}^\infty [\omega: \tau(\omega) = n, f(X_n(\omega)) = x]$ lies in $\mathcal{F}$, and so $f(X_\tau)$ is a simple random variable. In order that this always hold, put $f(X_{\tau(\omega)}(\omega)) = 0$, say, if $\tau(\omega) = \infty$ (which happens only on a set of probability 0).

The game with payoff function f has at i the *value*

$$(8.44) \qquad v(i) = \sup E_i[f(X_\tau)],$$

the supremum extending over all Markov times τ. It will turn out that the supremum here is achieved: there always exists an optimal stopping time. It will also turn out that there is an optimal τ that works for all initial states i. The problem is to calculate $v(i)$ and find the best τ. If the chain is irreducible, the system must pass through every state, and the best strategy is obviously to wait until the system enters a state for which f is maximal. This describes an optimal τ, and $v(i) = \max f$ for all i. For this reason the interesting cases are those in which some states are transient and others are absorbing ($p_{ii} = 1$).

A function φ on S is *excessive* or *superharmonic*, if[†]

$$(8.45) \qquad \varphi(i) \geq \sum_j p_{ij}\varphi(j), \qquad i \in S.$$

In terms of conditional expectation the requirement is $\varphi(i) \geq E_i[\varphi(X_1)]$.

Lemma 4. *The value function v is excessive.*

Proof. Given ϵ, choose for each j in S a "good" stopping time τ_j satisfying $E_j[f(X_{\tau_j})] > v(j) - \epsilon$. By (8.43), $[\tau_j = n] = [(X_0, \ldots, X_n) \in I_{jn}]$ for some set I_{jn} of $(n + 1)$-long sequences of states. Set $\tau = n + 1$ ($n \geq 0$) on the set $[X_1 = j] \cap [(X_1, \ldots, X_{n+1}) \in I_{jn}]$; that is, take one step and then from the new state X_1 add on the "good" stopping time for that state. Then τ is a

---

[†]Compare the conditions (7.28) and (7.35).

stopping time and

$$E_i[f(X_\tau)] = \sum_{n=0}^{\infty} \sum_j \sum_k P_i[X_1 = j, (X_1, \ldots, X_{n+1}) \in I_{jn}, X_{n+1} = k] f(k)$$

$$= \sum_{n=0}^{\infty} \sum_j \sum_k p_{ij} P_j[(X_0, \ldots, X_n) \in I_{jn}, X_n = k] f(k)$$

$$= \sum_j p_{ij} E_j[f(X_{\tau_j})].$$

Therefore, $v(i) \geq E_i[f(X_\tau)] \geq \sum_j p_{ij}(v(j) - \epsilon) = \sum_j p_{ij} v(j) - \epsilon$. Since ϵ was arbitrary, v is excessive.   ∎

Lemma 5. *Suppose that φ is excessive.*

(i) *For all stopping times τ, $\varphi(i) \geq E_i[\varphi(X_\tau)]$.*

(ii) *For all pairs of stopping times satisfying $\sigma \leq \tau$, $E_i[\varphi(X_\sigma)] \geq E_i[\varphi(X_\tau)]$.*

Part (i) says that for an excessive payoff function, $\tau \equiv 0$ represents an optimal strategy.

Proof.   To prove (i), put $\tau_N = \min\{\tau, N\}$. Then τ_N is a stopping time, and

$$(8.46) \qquad E_i[\varphi(X_{\tau_N})] = \sum_{n=0}^{N-1} \sum_k P_i[\tau = n, X_n = k]\varphi(k)$$

$$+ \sum_k P_i[\tau \geq N, X_N = k]\varphi(k).$$

Since $[\tau \geq N] = [\tau < N]^c \in F_{N-1}$, the final sum here is by (8.13)

$$\sum_k \sum_j P_i[\tau \geq N, X_{N-1} = j, X_N = k]\varphi(k)$$

$$= \sum_k \sum_j P_i[\tau \geq N, X_{N-1} = j] p_{jk}\varphi(k) \leq \sum_j P_i[\tau \geq N, X_{N-1} = j]\varphi(j).$$

Substituting this into (8.46) leads to $E_i[\varphi(X_{\tau_N})] \leq E_i[\varphi(X_{\tau_{N-1}})]$. Since $\tau_0 = 0$ and $E_i[\varphi(X_0)] = \varphi(i)$, it follows that $E_i[\varphi(X_{\tau_N})] \leq \varphi(i)$ for all N. But for $\tau(\omega)$ finite, $\varphi(X_{\tau_N(\omega)}(\omega)) \to \varphi(X_{\tau(\omega)}(\omega))$ (there is equality for large N), and so $E_i[\varphi(X_{\tau_N})] \to E_i[\varphi(X_\tau)]$ by Theorem 5.4.

The proof of (ii) is essentially the same. If $\tau_N = \min\{\tau, \sigma + N\}$, then τ_N is a stopping time, and

$$
E_i\left[\varphi(X_{\tau_N})\right] = \sum_{m=0}^{\infty} \sum_{n=0}^{N-1} \sum_k P_i[\sigma = m, \tau = m + n, X_{m+n} = k]\varphi(k)
$$

$$
+ \sum_{m=0}^{\infty} \sum_k P_i[\sigma = m, \tau \geq m + N, X_{m+N} = k]\varphi(k).
$$

Since $[\sigma = m, \tau \geq m + N] = [\sigma = m] - [\sigma = m, \tau < m + N] \in \mathscr{F}_{m+N-1}$, again $E_i[\varphi(X_{\tau_N})] \leq E_i[\varphi(X_{\tau_{N-1}})] \leq E_i[\varphi(X_{\tau_0})]$. Since $\tau_0 = \sigma$, part (ii) follows from part (i) by another passage to the limit.   ∎

Lemma 6. *If an excessive function φ dominates the payoff function f, then it dominates the value function v as well.*

By definition, to say that g dominates h is to say that $g(i) \geq h(i)$ for all i.

PROOF. By Lemma 5, $\varphi(i) \geq E_i[\varphi(X_\tau)] \geq E_i[f(X_\tau)]$ for all Markov times τ, and so $\varphi(i) \geq v(i)$ for all i.   ∎

Since $\tau \equiv 0$ is a stopping time, v dominates f. Lemmas 4 and 6 immediately characterize v:

Theorem 8.10. *The value function v is the minimal excessive function dominating f.*

There remains the problem of constructing the optimal strategy τ. Let M be the set of states i for which $v(i) = f(i)$; M, the *support set*, is nonempty, since it at least contains those i that maximize f. Let $A = \bigcap_{n=0}^{\infty}[X_n \notin M]$ be the event that the system never enters M. The following argument shows that $P_i(A) = 0$ for each i. As this is trivial if $M = S$, assume that $M \neq S$. Choose $\delta > 0$ so that $f(i) \leq v(i) - \delta$ for $i \in S - M$. Now $E_i[f(X_\tau)] = \sum_{n=0}^{\infty} \sum_k P_i[\tau = n, X_n = k]f(k)$; replacing the $f(k)$ by $v(k)$ or $v(k) - \delta$ according as $k \in M$ or $k \in S - M$ gives $E_i[f(X_\tau)] \leq E_i[v(X_\tau)] - \delta P_i[X_\tau \in S - M] \leq E_i[v(X_\tau)] - \delta P_i(A) \leq v(i) - \delta P_i(A)$, the last inequality by Lemmas 4 and 5. Since this holds for every Markov time, taking the supremum over τ gives $P_i(A) = 0$. Whatever the initial state, the system is thus certain to enter the support set M.

Let $\tau_0(\omega) = \min[n: X_n(\omega) \in M]$ be the *hitting time* for M. Then τ_0 is a Markov time, and $\tau_0 = 0$ if $X_0 \in M$. It may be that $X_n(\omega) \notin M$ for all n, in which case $\tau_0(\omega) = \infty$, but as just shown, the probability of this is 0.

Theorem 8.11.   *The hitting time τ_0 is optimal:* $E_i[f(X_{\tau_0})] = v(i)$ *for all* i.

PROOF.   By the definition of τ_0, $f(X_{\tau_0}) = v(X_{\tau_0})$. Put $\varphi(i) = E_i[f(X_{\tau_0})] = E_i[v(X_{\tau_0})]$. The first step is to show that φ is excessive. If $\tau_1 = \min[n: n \geq 1, X_n \in M]$, then τ_1 is a Markov time and

$$E_i\Big[v(X_{\tau_1})\Big] = \sum_{n=1}^{\infty} \sum_{k \in M} P_i[X_1 \notin M, \ldots, X_{n-1} \notin M, X_n = k]v(k)$$

$$= \sum_{n=1}^{\infty} \sum_{k \in M} \sum_{j \in S} p_{ij}P_j[X_0 \notin M, \ldots, X_{n-2} \notin M, X_{n-1} = k]v(k)$$

$$= \sum_{j} p_{ij}E_j\Big[v(X_{\tau_0})\Big].$$

Since $\tau_0 \leq \tau_1$, $E_i[v(X_{\tau_0})] \geq E_i[v(X_{\tau_1})]$ by Lemmas 4 and 5.

This shows that φ is excessive. And $\varphi(i) \leq v(i)$ by the definition (8.44). If $\varphi(i) \geq f(i)$ is proved, it will follow by Theorem 8.10 that $\varphi(i) \geq v(i)$ and hence that $\varphi(i) = v(i)$. Since $\tau_0 = 0$ for $X_0 \in M$, if $i \in M$ then $\varphi(i) = E_i[f(X_0)] = f(i)$. Suppose that $\varphi(i) < f(i)$ for some values of i in $S - M$, and choose i_0 to maximize $f(i) - \varphi(i)$. Then $\psi(i) = \varphi(i) + f(i_0) - \varphi(i_0)$ dominates f and is excessive, being the sum of a constant and an excessive function. By Theorem 8.10, ψ must dominate v, so that $\psi(i_0) \geq v(i_0)$, or $f(i_0) \geq v(v_0)$. But this implies that $i_0 \in M$, a contradiction   ∎

The optimal strategy need not be unique. If f is constant, for example, all strategies have the same value.

Example 8.16.   For the symmetric random walk with absorbing barriers at 0 and r (Example 8.2) a function φ on $S = \{0, 1, \ldots, r\}$ is excessive if $\varphi(i) \geq \frac{1}{2}\varphi(i - 1) + \frac{1}{2}\varphi(i + 1)$ for $1 \leq i \leq r - 1$. The requirement is that φ give a concave function when extended by linear interpolation from S to the entire interval $[0, r]$. Hence v thus extended is the minimal concave function dominating f. The figure shows the geometry: the ordinates of the dots are the values of f and the polygonal line describes v. The optimal strategy is to stop at a state for which the dot lies on the polygon.
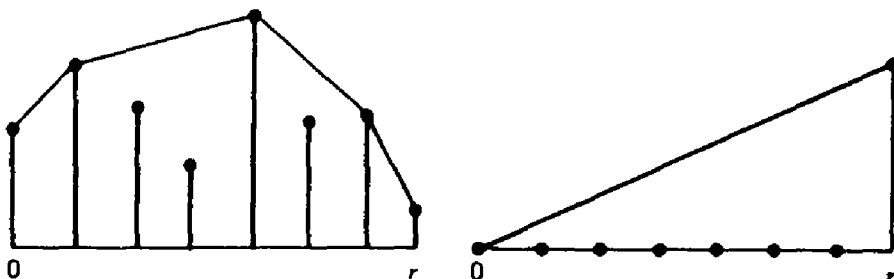
If $f(r) = 1$ and $f(i) = 0$ for $i < r$, then v is a straight line; $v(i) = i/r$. The optimal Markov time τ_0 is the hitting time for $M = \{0, r\}$, and $v(i) = E_i[f(X_{\tau_0})]$ is the probability of absorption in the state r. This gives another solution of the gambler's ruin problem for the symmetric case.   ■

Example 8.17.   For the selection problem in Example 8.5, the p_{ij} are given by (8.5) and (8.6) for $1 \le i \le r$, while $p_{r+1, r+1} = 1$. The payoff is $f(i) = i/r$ for $i \le r$ and $f(r + 1) = 0$. Thus $v(r + 1) = 0$, and since v is excessive,

$$(8.47) \qquad v(i) \ge g(i) = \sum_{j=i+1}^{r} \frac{i}{j(j+1)} v(j), \qquad 1 \le i < r.$$

By Theorem 8.10, v is the smallest function satisfying (8.47) and $v(i) \ge f(i) = i/r$, $1 \le i \le r$. Since (8.47) puts no lower limit on $v(r)$, it follows that $v(r) = f(r) = 1$, and r lies in the support set M. By minimality,

$$(8.48) \qquad v(i) = \max\{f(i), g(i)\}, \qquad 1 \le i < r.$$

If $i \in M$, then $f(i) = v(i) \ge g(i) \ge \sum_{j=i+1}^{r} ij^{-1}(j-1)^{-1}f(j) = f(i)\sum_{j=i+1}^{r}(j-1)^{-1}$, and hence $\sum_{j=i+1}^{r}(j-1)^{-1} \le 1$. On the other hand, if this inequality holds and $i + 1, \ldots, r$ all lie in M, then $g(i) = \sum_{j=i+1}^{r} ij^{-1}(j-1)^{-1}f(j) = f(i)\sum_{j=i+1}^{r}(j-1)^{-1} \le f(i)$, so that $i \in M$ by (8.48). Therefore, $M = \{i_r, i_r + 1, \ldots, r, r + 1\}$, where i_r is determined by

$$(8.49) \qquad \frac{1}{i_r} + \frac{1}{i_r + 1} + \cdots + \frac{1}{r-1} \le 1 < \frac{1}{i_r - 1} + \frac{1}{i_r} + \cdots + \frac{1}{r-1}$$

If $i < i_r$, so that $i \notin M$, then $v(i) > f(i)$ and so, by (8.48),

$$v(i) = g(i) = \sum_{j=i+1}^{i_r - 1} \frac{i}{j(j-1)} v(j) + \sum_{j=i_r}^{r} \frac{i}{j(j-1)} f(j)$$

$$= \sum_{j=i+1}^{i_r - 1} \frac{i}{j(j-1)} v(j) + \frac{i}{r} \left(\frac{1}{i_r - 1} + \cdots + \frac{1}{r-1} \right).$$

It follows by backward induction starting with $i = i_r - 1$ that

$$(8.50) \qquad v(i) = p_r = \frac{i_r - 1}{r} \left(\frac{1}{i_r - 1} + \cdots + \frac{1}{r-1} \right)$$

is constant for $1 \le i < i_r$.

In the selection problem as originally posed, $X_1 = 1$. The optimal strategy is to stop with the first X_n that lies in M. The princess should therefore reject the first $i_r - 1$ suitors and accept the next one who is preferable to all his predecessors (is dominant). The probability of success is p_r as given by (8.50). Failure can happen in two ways. Perhaps the first dominant suitor after i_r is not the best of all suitors; in this case the princess will be unaware of failure. Perhaps no dominant suitor comes after i_r; in this case the princess is obliged to take the last suitor of all and may be well

aware of failure. Recall that the problem was to maximize the chance of getting the best suitor of all rather than, say, the chance of getting a suitor in the top half.

If r is large, (8.49) essentially requires that $\log r - \log i_r$ be near 1, so that $i_r \approx r/e$. In this case, $p_r \approx 1/e$.

Note that although the system starts in state 1 in the original problem, its resolution by means of the preceding theory requires consideration of all possible initial states.   ∎

This theory carries over in part to the case of infinite S, although this requires the general theory of expected values, since $f(X_r)$ may not be a simple random variable. Theorem 8.10 holds for infinite S if the payoff function is nonnegative and the value function is finite.[†] But then problems arise: Optimal strategies may not exist, and the probability of hitting the support set M may be less than 1. Even if this probability is 1, the strategy of stopping on first entering M may be the worst one of all.[‡]

PROBLEMS

8.1. Prove Theorem 8.1 for the case of finite S by constructing the appropriate probability measure on sequence space S^∞: Replace the summand on the right in (2.21) by $\alpha_{u_1} p_{u_1 u_2} \cdots p_{u_{n-1}u_n}$, and extend the arguments preceding Theorem 2.3. If $X_n(\cdot) = z_n(\cdot)$, then $X_1, X_2, \ldots$ is the appropriate Markov chain (here time is shifted by 1).

8.2. Let $Y_0, Y_1, \ldots$ be independent and identically distributed with $P[Y_n = 1] = p$, $P[Y_n = 0] = q = 1 - p, p \neq q$. Put $X_n = Y_n + Y_{n+1}$ (mod 2). Show that $X_0, X_1, \ldots$ is not a Markov chain even though $P[X_{n+1} = j | X_{n-1} = i] = P[X_{n+1} = j]$. Does this last relation hold for all Markov chains? Why?

8.3. Show by example that a function $f(X_0), f(X_1), \ldots$ of a Markov chain need not be a Markov chain.

8.4. Show that

$$f_{ij} \sum_{k=0}^{\infty} p_{jj}^{(k)} = \sum_{n=1}^{\infty} \sum_{m=1}^{n} f_{ij}^{(m)} p_{jj}^{(n-m)} = \sum_{n=1}^{\infty} p_{ij}^{(n)},$$

and prove that if j is transient, then $\sum_n p_{ij}^{(n)} < \infty$ for each i (compare Theorem 8.3(i)). If j is transient, then

$$f_{ij} = \sum_{n=1}^{\infty} p_{ij}^{(n)} \bigg/ \left(1 + \sum_{n=1}^{\infty} p_{jj}^{(n)}\right).$$

[†]The only essential change in the argument is that Fatou's lemma (Theorem 16.3) must be used in place of Theorem 5 4 in the proof of Lemma 5.
[‡]See Problems 8 36 and 8.37

Specialize to the case $i = j$: in addition to implying that i is transient (Theorem 8.2(i)), a finite value for $\sum_{n=1}^{\infty} p_{ii}^{(n)}$ suffices to determine f_{ii} exactly.

8.5. Call $\{x_i\}$ a *subsolution* of (8.24) if $x_i \le \sum_j q_{ij} x_j$ and $0 \le x_i \le 1$, $i \in U$. Extending Lemma 1, show that a subsolution $\{x_i\}$ satisfies $x_i \le \sigma_i$: The solution $\{\sigma_i\}$ of (8.24) dominates all subsolutions as well as all solutions. Show that if $x_i = \sum_j q_{ij} x_j$ and $-1 \le x_i \le 1$, then $\{|x_i|\}$ is a subsolution of (8.24).

8.6. Show by solving (8.27) that the unrestricted random walk on the line (Example 8.3) is persistent if and only if $p = \frac{1}{2}$

8.7. (a) Generalize an argument in the proof of Theorem 8.5 to show that $f_{ik} = p_{ik} + \sum_{j \ne k} p_{ij} f_{jk}$. Generalize this further to

$$f_{ik} = f_{ik}^{(1)} + \cdots + f_{ik}^{(n)}$$
$$+ \sum_{j \ne k} P_i[X_1 \ne k, \ldots, X_{n-1} \ne k, X_n = j] f_{jk}$$

(b) Take $k = i$. Show that $f_{ij} > 0$ if and only if $P_i[X_1 \ne i, \ldots, X_{n-1} \ne i, X_n = j] > 0$ for some n, and conclude that i is transient if and only if $f_{ji} < 1$ for some $j \ne i$ such that $f_{ij} > 0$.

(c) Show that an irreducible chain is transient if and only if for each i there is a $j \ne i$ such that $f_{ji} < 1$.

8.8. Suppose that $S = \{0, 1, 2, \ldots\}$, $p_{00} = 1$, and $f_{i0} > 0$ for all i.

(a) Show that $P_i(\bigcup_{j=1}^{\infty} [X_n = j \text{ i.o.}]) = 0$ for all i.

(b) Regard the state as the size of a population and interpret the conditions $p_{00} = 1$ and $f_{i0} > 0$ and the conclusion in part (a).

8.9. 8.5↑   Show for an irreducible chain that (8.27) has a nontrivial solution if and only if there exists a nontrivial, bounded sequence $\{x_i\}$ (not necessarily nonnegative) satisfying $x_i = \sum_{j \ne i_0} p_{ij} x_j$, $i \ne i_0$. (See the remark following the proof of Theorem 8.5.)

8.10. ↑   Show that an irreducible chain is transient if and only if (for arbitrary i_0) the system $y_i = \sum_j p_{ij} y_j$, $i \ne i_0$ (sum over *all* j), has a bounded, nonconstant solution $\{y_i, i \in S\}$.

8.11. Show that the P_i-probabilities of ever leaving U for $i \in U$ are the minimal solution of the system.

$$(8.51) \qquad \begin{cases} z_i = \sum_{j \in U} p_{ij} z_j + \sum_{j \notin U} p_{ij}, & i \in U, \\ 0 \le z_i \le 1, & i \in U. \end{cases}$$

The constraint $z_i \le 1$ can be dropped: the minimal solution automatically satisfies it, since $z_i \equiv 1$ is a solution.

8.12. Show that $\sup_{ij} n_0(i, j) = \infty$ is possible in Lemma 2.

8.13. Suppose that $\{\pi_i\}$ solves (8.30), where it is assumed that $\sum_i |\pi_i| < \infty$, so that the left side is well defined. Show in the irreducible case that the π_i are either all positive or all negative or all 0. Stationary probabilities thus exist in the irreducible case if and only if (8.30) has a nontrivial solution $\{\pi_i\}$ ($\sum_i \pi_i$ absolutely convergent).

8.14. Show by example that the coupled chain in the proof of Theorem 8.6 need not be irreducible if the original chain is not aperiodic.

8.15. Suppose that S consists of all the integers and

$$p_{0,-1} = p_{0,0} = p_{0,+1} = \tfrac{1}{3},$$

$$p_{k,k-1} = q, \quad p_{k,k+1} = p, \qquad k \le -1,$$

$$p_{k,k-1} = p, \quad p_{k,k+1} = q, \qquad k \ge 1.$$

Show that the chain is irreducible and aperiodic. For which p's is the chain persistent? For which p's are there stationary probabilities?

8.16. Show that the period of j is the greatest common divisor of the set

(8.52) $$\left[n\colon n \ge 1, f_{jj}^{(n)} > 0 \right].$$

8.17. $\uparrow$ *Recurrent events.* Let $f_1, f_2, \ldots$ be nonnegative numbers with $f = \sum_{n=1}^{\infty} f_n \le 1$. Define $u_1, u_2, \ldots$ recursively by $u_1 = f_1$ and

(8.53) $$u_n = f_1 u_{n-1} + \cdots + f_{n-1} u_1 + f_n.$$

(a) Show that $f < 1$ if and only if $\sum_n u_n < \infty$.
(b) Assume that $f = 1$, set $\mu = \sum_{n=1}^{\infty} n f_n$, and assume that

(8.54) $$\gcd[n\colon n \ge 1, f_n > 0] = 1.$$

Prove the *renewal theorem*. Under these assumptions, the limit $u = \lim_n u_n$ exists, and $u > 0$ if and only if $\mu < \infty$, in which case $u = 1/\mu$.

Although these definitions and facts are stated in purely analytical terms, they have a probabilistic interpretation: Imagine an event $\mathscr{E}$ that may occur at times $1, 2, \ldots$. Suppose f_n is the probability $\mathscr{E}$ occurs first at time n. Suppose further that at each occurrence of $\mathscr{E}$ the system starts anew, so that f_n is the probability that $\mathscr{E}$ next occurs n steps later. Such an $\mathscr{E}$ is called a *recurrent event.* If u_n is the probability that $\mathscr{E}$ occurs at time n, then (8.53) holds. The recurrent event $\mathscr{E}$ is called transient or persistent according as $f < 1$ or $f = 1$, it is called aperiodic if (8.54) holds, and if $f = 1$, μ is interpreted as the mean recurrence time

8.18. (a) Let τ be the smallest integer for which $X_\tau = i_0$. Suppose that the state space is finite and that the p_{ij} are all positive. Find a ρ such that $\max_i (1 - p_{i i_0}) \le \rho < 1$ and hence $P_i[\tau > n] \le \rho^n$ for all i.
(b) Apply this to the coupled chain in the proof of Theorem 8.6: $|p_{ik}^{(n)} - p_{jk}^{(n)}| \le \rho^n$. Now give a new proof of Theorem 8.9.

8.19. A thinker who owns r umbrellas wanders back and forth between home and office, taking along an umbrella (if there is one at hand) in rain (probability p) but not in shine (probability q). Let the state be the number of umbrellas at hand, irrespective of whether the thinker is at home or at work. Set up the transition matrix and find the stationary probabilities. Find the steady-state probability of his getting wet, and show that five umbrellas will protect him at the 5% level against any climate (any p).

8.20. (a) A transition matrix is *doubly stochastic* if $\sum_i p_{ij} = 1$ for each j. For a finite, irreducible, aperiodic chain with doubly stochastic transition matrix, show that the stationary probabilities are all equal.

(b) Generalize Example 8.15: Let S be a finite group, let $p(i)$ be probabilities, and put $p_{ij} = p(j \cdot i^{-1})$, where product and inverse refer to the group operation. Show that, if all $p(i)$ are positive, the states are all equally likely in the limit.

(c) Let S be the symmetric group on 52 elements. What has (b) to say about card shuffling?

8.21. A set C in S is *closed* if $\sum_{j \in C} p_{ij} = 1$ for $i \in C$: once the system enters C it cannot leave. Show that a chain is irreducible if and only if S has no proper closed subset.

8.22. ↑ Let T be the set of transient states and define persistent states i and j (if there are any) to be equivalent if $f_{ij} > 0$. Show that this is an equivalence relation on $S - T$ and decomposes it into equivalence classes $C_1, C_2, \ldots$, so that $S = T \cup C_1 \cup C_2 \cup \cdots$ Show that each C_m is closed and that $f_{ij} = 1$ for i and j in the same C_m.

8.23. 8.11 8.21 ↑ Let T be the set of transient states and let C be any closed set of persistent states. Show that the P_i-probabilities of eventual absorption in C for $i \in T$ are the minimal solution of

$$(8.55) \qquad \begin{cases} y_i = \sum_{j \in T} p_{ij} y_j + \sum_{j \in C} p_{ij}, & i \in T, \\ 0 \le y_i \le 1, & i \in T. \end{cases}$$

8.24. Suppose that an irreducible chain has period $t > 1$. Show that S decomposes into sets $S_0, \ldots, S_{t-1}$ such that $p_{ij} > 0$ only if $i \in S_\nu$ and $j \in S_{\nu+1}$ for some ν ($\nu + 1$ reduced modulo t). Thus the system passes through the S_ν in cyclic succession.

8.25. ↑ Suppose that an irreducible chain of period $t > 1$ has a stationary distribution $\{\pi_j\}$. Show that, if $i \in S_\nu$ and $j \in S_{\nu + \alpha}$ ($\nu + \alpha$ reduced modulo t), then $\lim_n p_{ij}^{(nt + \alpha)} = \pi_j$. Show that $\lim_n n^{-1} \sum_{m=1}^n p_{ij}^{(m)} = \pi_j / t$ for all i and j.

8.26. *Eigenvalues.* Consider an irreducible, aperiodic chain with state space $\{1, \ldots, s\}$. Let $r_0 = (\pi_1, \ldots, \pi_s)$ be (Example 8.14) the row vector of stationary probabilities, and let c_0 be the column vector of 1's; then r_0 and c_0 are left and right eigenvectors of P for the eigenvalue $\lambda = 1$.

(a) Suppose that r is a left eigenvector for the (possibly complex) eigenvalue λ: $rP = \lambda r$. Prove: If $\lambda = 1$, then r is a scalar multiple of r_0 ($\lambda = 1$ has geometric

multiplicity 1). If $\lambda \neq 1$, then $|\lambda| < 1$ and $rc_0 = 0$ (the 1×1 product of $1 \times s$ and $s \times 1$ matrices).

(b) Suppose that c is a right eigenvector: $Pc = \lambda c$. If $\lambda = 1$, then c is a scalar multiple of c_0 (again the geometric multiplicity is 1). If $\lambda \neq 1$, then again $|\lambda| < 1$, and $r_0 c = 0$.

8.27. ↑ Suppose P is diagonalizable; that is, suppose there is a nonsingular C such that $C^{-1}PC = \Lambda$, where Λ is a diagonal matrix. Let $\lambda_1, \ldots, \lambda_s$ be the diagonal elements of Λ, let $c_1, \ldots, c_s$ be the successive columns of C, let $R = C^{-1}$, and let $r_1, \ldots, r_s$ be the successive rows of R.

(a) Show that c_i and r_i are right and left eigenvectors for the eigenvalue λ_i, $i = 1, \ldots, s$. Show that $r_i c_j = \delta_{ij}$. Let $A_i = c_i r_i$ ($s \times s$). Show that Λ^n is a diagonal matrix with diagonal elements $\lambda_1^n, \ldots, \lambda_s^n$ and that $P^n = C\Lambda^n R = \sum_{u=1}^{s} \lambda_u^n A_u, n \geq 1$.

(b) Part (a) goes through under the sole assumption that P is a diagonalizable matrix. Now assume also that it is an irreducible, aperiodic stochastic matrix, and arrange the notation so that $\lambda_1 = 1$. Show that each row of A_1 is the vector $(\pi_1, \ldots, \pi_s)$ of stationary probabilities. Since

$$(8.56) \qquad\qquad P^n = A_1 + \sum_{u=2}^{s} \lambda_u^n A_u$$

and $|\lambda_u| < 1$ for $2 \leq u \leq s$, this proves exponential convergence once more.

(c) Write out (8.56) explicitly for the case $s = 2$.

(d) Find an irreducible, aperiodic stochastic matrix that is not diagonalizable.

8.28. ↑ (a) Show that the eigenvalue $\lambda = 1$ has geometric multiplicity 1 if there is only one closed, irreducible set of states; there may be transient states, in which case the chain itself is not irreducible.

(b) Show, on the other hand, that if there is more than one closed, irreducible set of states, then $\lambda = 1$ has geometric multiplicity exceeding 1.

(c) Suppose that there is only one closed, irreducible set of states. Show that the chain has period exceeding 1 if and only if there is an eigenvalue other than 1 on the unit circle.

8.29. Suppose that $\{X_n\}$ is a Markov chain with state space S, and put $Y_n = (X_n, X_{n+1})$. Let T be the set of pairs (i, j) such that $p_{ij} > 0$ and show that $\{Y_n\}$ is a Markov chain with state space T. Write down the transition probabilities. Show that, if $\{X_n\}$ is irreducible and aperiodic, so is $\{Y_n\}$. Show that, if π_i are stationary probabilities for $\{X_n\}$, then $\pi_i p_{ij}$ are stationary probabilities for $\{Y_n\}$.

8.30. 6.10 8.29 ↑ Suppose that the chain is finite, irreducible, and aperiodic and that the initial probabilities are the stationary ones. Fix a state i, let $A_n = [X_i = i]$, and let N_n be the number of passages through i in the first n steps. Calculate α_n and β_n as defined by (5.41). Show that $\beta_n - \alpha_n^2 = O(1/n)$, so that $n^{-1}N_n \to \pi_i$ with probability 1. Show for a function f on the state space that $n^{-1}\sum_{k=1}^{n} f(X_k) \to \sum_i \pi_i f(i)$ with probability 1. Show that $n^{-1}\sum_{k=1}^{n} g(X_k, X_{k+1}) \to \sum_{ij} \pi_i p_{ij} g(i, j)$ for functions g on $S \times S$.

8.31. 6.14 8.30↑ If $X_0(\omega) = i_0, \ldots, X_n(\omega) = i_n$ for states $i_0, \ldots, i_n$, put $p_n(\omega) = \pi_{i_0} p_{i_0 i_1} \cdots p_{i_{n-1} i_n}$, so that $p_n(\omega)$ is the probability of the observation observed. Show that $-n^{-1} \log p_n(\omega) \to h = -\sum_{ij} \pi_i p_{ij} \log p_{ij}$ with probability 1 if the chain is finite, irreducible, and aperiodic. Extend to this case the notions of source, entropy, and asymptotic equipartition.

8.32. A sequence $\{X_n\}$ is a Markov chain *of second order* if $P[X_{n+1} = j | X_0 = i_0, \ldots, X_n = i_n] = P[X_{n+1} = j | X_{n-1} = i_{n-1}, X_n = i_n] = p_{i_{n-1} i_n j}$. Show that nothing really new is involved because the sequence of pairs (X_n, X_{n+1}) is an ordinary Markov chain (of first order). Compare Problem 8.29. Generalize this idea into chains of order r.

8.33. Consider a chain on $S = \{0, 1, \ldots, r\}$, where 0 and r are absorbing states and $p_{i, i+1} = p_i > 0$, $p_{i, i-1} = q_i = 1 - p_i > 0$ for $0 < i < r$. Identify state i with a point z_i on the line, where $0 = z_0 < \cdots < z_r$ and the distance from z_i to z_{i+1} is q_i/p_i times that from z_{i-1} to z_i. Given a function φ on S, consider the associated function $\hat{\varphi}$ on $[0, z_r]$ defined at the z_i by $\hat{\varphi}(z_i) = \varphi(i)$ and in between by linear interpolation. Show that φ is excessive if and only if $\hat{\varphi}$ is concave. Show that the probability of absorption in r for initial state i is t_{i-1}/t_{r-1}, where $t_i = \sum_{k=0}^{i} q_1 \cdots q_k / p_1 \cdots p_k$. Deduce (7.7). Show that in the new scale the expected distance moved on each step is 0.

8.34. Suppose that a finite chain is irreducible and aperiodic. Show by Theorem 8.9 that an excessive function must be constant.

8.35. *A zero-one law.* Let the state space S contain s points, and suppose that $\epsilon_n = \sup_{ij} |p_{ij}^{(n)} - \pi_j| \to 0$, as holds under the hypotheses of Theorem 8.9. For $a \le b$, let $\mathcal{S}_a^b$ be the σ-field generated by the sets $[X_a = u_a, \ldots, X_b = u_b]$. Let $\mathcal{T}_a = \sigma(\bigcup_{b=a}^{\infty} \mathcal{S}_a^b)$ and $\mathcal{T} = \bigcap_{a=1}^{\infty} \mathcal{T}_a$. Show that $|P(A \cap B) - P(A)P(B)| \le s(\epsilon_n + \epsilon_{b+n})$ for $A \in \mathcal{S}_0^b$ and $B \in \mathcal{S}_{b+n}^{b+m}$; the ϵ_{b+n} can be suppressed if the initial probabilities are the stationary ones. Show that this holds for $A \in \mathcal{S}_0^b$ and $B \in \mathcal{T}_{b+n}$. Show that $C \in \mathcal{T}$ implies that $P(C)$ is either 0 or 1.

8.36† Alter the chain in Example 8.13 so that $q_0 = 1 - p_0 = 1$ (the other p_i and q_i still positive). Let $\beta = \lim_n p_1 \cdots p_n$ and assume that $\beta > 0$. Define a payoff function by $f(0) = 1$ and $f(i) = 1 - f_{i0}$ for $i > 0$. If $X_0, \ldots, X_n$ are positive, put $\sigma_n = n$; otherwise let σ_n be the smallest k such that $X_k = 0$. Show that $E_i[f(X_{\sigma_n})] \to 1$ as $n \to \infty$, so that $v(i) \equiv 1$. Thus the support set is $M = \{0\}$, and for an initial state $i > 0$ the probability of ever hitting M is $f_{i0} < 1$.

For an arbitrary finite stopping time τ, choose n so that $P_i[\tau < n = \sigma_n] > 0$. Then $E_i[f(X_\tau)] \le 1 - f_{i+n,0} P_i[\tau < n = \sigma_n] < 1$. Thus no strategy achieves the value $v(i)$ (except of course for $i = 0$).

8.37. ↑ Let the chain be as in the preceding problem, but assume that $\beta = 0$, so that $f_{i0} = 1$ for all i. Suppose that $\lambda_1, \lambda_2, \ldots$ exceed 1 and that $\lambda_1 \cdots \lambda_n \to \lambda < \infty$; put $f(0) = 0$ and $f(i) = \lambda_1 \cdots \lambda_{i-1}/p_1 \cdots p_{i-1}$. For an arbitrary (finite) stopping time τ, the event $[\tau = n]$ must have the form $[(X_0, \ldots, X_n) \in I_n]$ for some set I_n of $(n+1)$-long sequences of states. Show that for each i there is at

---

†The final three problems in this section involve expected values for random variables with infinite range.

most one $n \geq 0$ such that $(i, i+1, \ldots, i+n) \in I_n$. If there is no such n, then $E_i[f(X_\tau)] = 0$. If there is one, then

$$E_i[f(X_\tau)] = P_i[(X_0, \ldots, X_n) = (i, \ldots, i+n)]f(i+n),$$

and hence the only possible values of $E_i[f(X_\tau)]$ are

$$0, \quad f(i), \quad p_i f(i+1) = f(i)\lambda_i, \quad p_i p_{i+1} f(i+2) = f(i)\lambda_i \lambda_{i+1}, \ldots.$$

Thus $v(i) = f(i)\lambda/\lambda_1 \cdots \lambda_{i-1}$ for $i \geq 1$; no strategy this value. The support set is $M = \{0\}$, and the hitting time τ_0 for M is finite, but $E_i[f(X_{\tau_0})] = 0$.

8.38. 5.12↑   Consider an irreducible, aperiodic, positive persistent chain. Let τ_j be the smallest n such that $X_n = j$, and let $m_{ij} = E_i[\tau_j]$. Show that there is an r such that $p = P_j[X_1 \neq j, \ldots, X_{r-1} \neq j, X_r = i]$ is positive; from $f_{jj}^{(n+r)} \geq p f_{ij}^{(n)}$ and $m_{jj} < \infty$, conclude that $m_{ij} < \infty$ and $m_{ij} = \sum_{n=0}^{\infty} P_i[\tau_j > n]$. Starting from $p_{ij}^{(t)} = \sum_{s=1}^{t} f_{ij}^{(s)} p_{jj}^{(t-s)}$, show that

$$\sum_{t=1}^{n} \left(p_{ij}^{(t)} - p_{jj}^{(t)} \right) = 1 - \sum_{m=0}^{n} p_{jj}^{(n-m)} P_i[\tau_i > m].$$

Use the M-test to show that

$$\pi_j m_{ij} = 1 + \sum_{n=1}^{\infty} \left(p_{jj}^{(n)} - p_{ij}^{(n)} \right).$$

If $i = j$, this gives $m_{jj} = 1/\pi_j$ again; if $i \neq j$, it shows how in principle m_{ij} can be calculated from the transition matrix and the stationary probabilities.

SECTION 9. LARGE DEVIATIONS AND THE LAW OF THE ITERATED LOGARITHM*

It is interesting in connection with the strong law of large numbers to estimate the rate at which S_n/n converges to the mean m. The proof of the strong law used upper bounds for the probabilities $P[|S_n - m| \geq \alpha]$ for large α. Accurate upper and lower bounds for these probabilities will lead to the law of the iterated logarithm, a theorem giving very precise rates for $S_n/n \to m$.

The first concern will be to estimate the probability of large deviations from the mean, which will require the method of moment generating functions. The estimates will be applied first to a problem in statistics and then to the law of the iterated logarithm.

---

*This section may be omitted

Moment Generating Functions

Let X be a simple random variable asssuming the distinct values $x_1, \ldots, x_l$ with respective probabilities $p_1, \ldots, p_l$. Its *moment generating function* is

$$(9.1) \qquad M(t) = E[e^{tX}] = \sum_{i=1}^{l} p_i e^{tx_i}.$$

(See (5.19) for expected values of functions of random variables.) This function, defined for all real t, can be regarded as associated with X itself or as associated with its distribution—that is, with the measure on the line having mass p_i at x_i (see (5.12)).

If $c = \max_i |x_i|$, the partial sums of the series $e^{tX} = \sum_{k=0}^{\infty} t^k X^k / k!$ are bounded by $e^{|t|c}$, and so the corollary to Theorem 5.4 applies:

$$(9.2) \qquad M(t) = \sum_{k=0}^{\infty} \frac{t^k}{k!} E[X^k].$$

Thus $M(t)$ has a Taylor expansion, and as follows from the general theory [A29], the coefficient of t^k must be $M^{(k)}(0)/k!$ Thus

$$(9.3) \qquad E[X^k] = M^{(k)}(0).$$

Furthermore, term-by-term differentiation in (9.1) gives

$$M^{(k)}(t) = \sum_{i=1}^{l} p_i x_i^k e^{tx_i} = E[X^k e^{tX}];$$

taking $t = 0$ here gives (9.3) again. Thus the moments of X can be calculated by successive differentiation, whence $M(t)$ gets its name. Note that $M(0) = 1$.

Example 9.1. If X assumes the values 1 and 0 with probabilities p and $q = 1 - p$, as in Bernoulli trials, its moment generating function is $M(t) = pe^t + q$. The first two moments are $M'(0) = p$ and $M''(0) = p$, and the variance is $p - p^2 = pq$. ∎

If $X_1, \ldots, X_n$ are independent, then for each t (see the argument following (5.10)), $e^{tX_1}, \ldots, e^{tX_n}$ are also independent. Let M and $M_1, \ldots, M_n$ be the respective moment generating functions of $S = X_1 + \cdots + X_n$ and of $X_1, \ldots, X_n$; of course, $e^{tS} = \prod_i e^{tX_i}$. Since by (5.25) expected values multiply for independent random variables, there results the fundamental relation

$$(9.4) \qquad M(t) = M_1(t) \cdots M_n(t).$$

This is an effective way of calculating the moment generating function of the sum S. The real interest, however, centers on the distribution of S, and so it is important to know that distributions can in principle be recovered from their moment generating functions.

Consider along with (9.1) another finite exponential sum $N(t) = \sum_j q_j e^{ty_j}$, and suppose that $M(t) = N(t)$ for all t. If $x_{i_0} = \max x_i$ and $y_{j_0} = \max y_j$, then $M(t) \sim p_{i_0} e^{tx_{i_0}}$ and $N(t) \sim q_{j_0} e^{ty_{j_0}}$ as $t \to \infty$, and so $x_{i_0} = y_{j_0}$ and $p_{i_0} = q_{i_0}$. The same argument now applies to $\sum_{i \neq i_0} p_i e^{tx_i} = \sum_{j \neq j_0} q_j e^{ty_j}$, and it follows inductively that with appropriate relabeling, $x_i = y_i$ and $p_i = q_i$ for each i. Thus *the function* (9.1) *does uniquely determine the* x_i *and* p_i.

Example 9.2. If $X_1, \ldots, X_n$ are independent, each assuming values 1 and 0 with probabilities p and q, then $S = X_1 + \cdots + X_n$ is the number of successes in n Bernoulli trials. By (9.4) and Example 9.1, S has the moment generating function

$$E[e^{tS}] = (pe^t + q)^n = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} e^{tk}.$$

The right-hand form shows this to be the moment generating function of a distribution with mass $\binom{n}{k} p^k q^{n-k}$ at the integer k, $0 \le k \le n$. The uniqueness just established therefore yields the standard fact that $P[S = k] = \binom{n}{k} p^k q^{n-k}$.
∎

The *cumulant generating function of* X (or of its distribution) is

(9.5) $C(t) = \log M(t) = \log E[e^{tX}].$

(Note that $M(t)$ is strictly positive.) Since $C' = M'/M$ and $C'' = (MM'' - (M')^2)/M^2$, and since $M(0) = 1$,

(9.6) $C(0) = 0,$ $C'(0) = E[X],$ $C''(0) = \mathrm{Var}[X].$

Let $m_k = E[X^k]$. The leading term in (9.2) is $m_0 = 1$, and so a formal expansion of the logarithm in (9.5) gives

(9.7) $C(t) = \sum_{v=1}^{\infty} \frac{(-1)^{v+1}}{v} \left(\sum_{k=1}^{\infty} \frac{m_k}{k!} t^k \right)^v.$

Since $M(t) \to 1$ as $t \to 0$, this expression is valid for t in some neighborhood of 0. By the theory of series, the powers on the right can be expanded and

terms with a common factor t^i collected together. This gives an expansion

$$(9.8) \qquad C(t) = \sum_{i=1}^{\infty} \frac{c_i}{i!} t^i,$$

valid in some neighborhood of 0.

The c_i are the *cumulants* of X. Equating coefficients in the expansions (9.7) and (9.8) leads to $c_1 = m_1$ and $c_2 = m_2 - m_1^2$, which checks with (9.6). Each c_i can be expressed as a polynomial in $m_1, \ldots, m_i$ and conversely, although the calculations soon become tedious. If $E[X] = 0$, however, so that $m_1 = c_1 = 0$, it is not hard to check that

$$(9.9) \qquad c_3 = m_3, \qquad c_4 = m_4 - 3m_2^2.$$

Taking logarithms converts the multiplicative relation (9.4) into the additive relation

$$(9.10) \qquad C(t) = C_1(t) + \cdots + C_n(t)$$

for the corresponding cumulant generating functions; it is valid in the presence of independence. By this and the definition (9.8), it follows that cumulants add for independent random variables.

Clearly, $M''(t) = E[X^2 e^{tX}] \geq 0$. Since $(M'(t))^2 = E^2[Xe^{tX}] \leq E[e^{tX}] \cdot E[X^2 e^{tX}] = M(t)M''(t)$ by Schwarz's inequality (5.36), $C''(t) \geq 0$. Thus *the moment generating function and the cumulant generating function are both convex.*

Large Deviations

Let Y be a simple random variable assuming values y_j with probabilities p_j. The problem is to estimate $P[Y \geq \alpha]$ when Y has mean 0 and α is positive. It is notationally convenient to subtract α away from Y and instead estimate $P[Y \geq 0]$ when Y has negative mean.

Assume then that

$$(9.11) \qquad E[Y] < 0, \qquad P[Y > 0] > 0,$$

the second assumption to avoid trivialities. Let $M(t) = \sum_j p_j e^{ty_j}$ be the moment generating function of Y. Then $M'(0) < 0$ by the first assumption in
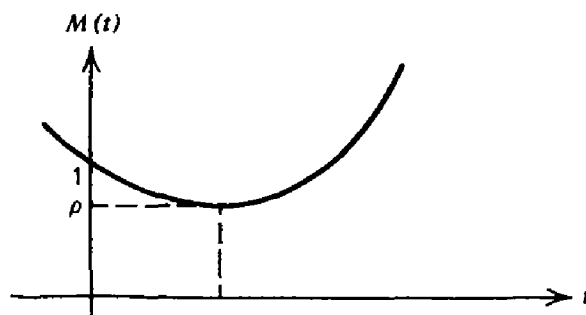
(9.11), and $M(t) \to \infty$ as $t \to \infty$ by the second. Since $M(t)$ is convex, it has its minimum ρ at a positive argument τ:

$$(9.12) \qquad \inf_t M(t) = M(\tau) = \rho, \qquad 0 < \rho < 1, \quad \tau > 0.$$

Construct (on an entirely irrelevant probability space) an auxiliary random variable Z such that

$$(9.13) \qquad P[Z = y_j] = \frac{e^{\tau y_j}}{\rho} P[Y = y_j]$$

for each y_j in the range of Y. Note that the probabilities on the right do add to 1. The moment generating function of Z is

$$(9.14) \qquad E[e^{tZ}] = \sum_j \frac{e^{\tau y_j}}{\rho} p_j e^{t y_j} = \frac{M(\tau + t)}{\rho},$$

and therefore

$$(9.15) \qquad E[Z] = \frac{M'(\tau)}{\rho} = 0, \qquad s^2 = E[Z^2] = \frac{M''(\tau)}{\rho} > 0.$$

For all positive t, $P[Y \geq 0] = P[e^{tY} \geq 1] \leq M(t)$ by Markov's inequality (5.31), and hence

$$(9.16) \qquad P[Y \geq 0] \leq \rho.$$

Inequalities in the other direction are harder to obtain. If $\sum'$ denotes summation over those indices j for which $y_j \geq 0$, then

$$(9.17) \qquad P[Y \geq 0] = \sum{}' p_j = \rho \sum{}' e^{-\tau y_j} P[Z = y_j].$$

Put the final sum here in the form $e^{-\theta}$, and let $p = P[Z \geq 0]$. By (9.16), $\theta \geq 0$. Since $\log x$ is concave, Jensen's inequality (5.33) gives

$$-\theta = \log \sum{}' e^{-\tau y_j} p^{-1} P[Z = y_j] + \log p$$

$$\geq \sum{}' (-\tau y_j) p^{-1} P[Z = y_j] + \log p$$

$$= -\tau s p^{-1} \sum{}' \frac{y_j}{s} P[Z = y_j] + \log p.$$

By (9.15) and Lyapounov's inequality (5.37),

$$\sum{}' \frac{y_j}{s} P[Z = y_j] \leq \frac{1}{s} E[|Z|] \leq \frac{1}{s} E^{1/2}[Z^2] = 1.$$

The last two inequalities give

$$(9.18) \qquad 0 \le \theta \le \frac{\tau s}{P[Z \ge 0]} - \log P[Z \ge 0].$$

This proves the following result.

Theorem 9.1. *Suppose that Y satisfies* (9.11). *Define* ρ *and* τ *by* (9.12), *let Z be a random variable with distribution* (9.13), *and define* s^2 *by* (9.15). *Then* $P[Y \ge 0] = \rho e^{-\theta}$, *where* θ *satisfies* (9.18).

To use (9.18) requires a lower bound for $P[Z \ge 0]$.

Theorem 9.2. *If* $E[Z] = 0$, $E[Z^2] = s^2$, *and* $E[Z^4] = \xi^4 > 0$, *then* $P[Z \ge 0] \ge s^4/4\xi^4$.[†]

PROOF. Let $Z^+ = ZI_{[Z \ge 0]}$ and $Z^- = -ZI_{[Z < 0]}$. Then Z^+ and Z^- are nonnegative, $Z = Z^+ - Z^-$, $Z^2 = (Z^+)^2 + (Z^-)^2$, and

$$(9.19) \qquad s^2 = E\left[(Z^+)^2\right] + E\left[(Z^-)^2\right].$$

Let $p = P[Z \ge 0]$. By Schwarz's inequality (5.36),

$$E\left[(Z^+)^2\right] = E\left[I_{[Z \ge 0]}Z^2\right]$$

$$\le E^{1/2}\left[I^2_{[Z \ge 0]}\right] E^{1/2}\left[Z^4\right] = p^{1/2}\xi^2.$$

By Hölder's inequality (5.35) (for $p = \frac{3}{2}$ and $q = 3$)

$$E\left[(Z^-)^2\right] = E\left[(Z^-)^{2/3}(Z^-)^{4/3}\right]$$

$$\le E^{2/3}[Z^-] E^{1/3}\left[(Z^-)^4\right] \le E^{2/3}[Z^-]\xi^{4/3}.$$

Since $E[Z] = 0$, another application of Hölder's inequality (for $p = 4$ and $q = \frac{4}{3}$) gives

$$E[Z^-] = E[Z^+] = E\left[ZI_{[Z \ge 0]}\right]$$

$$\le E^{1/4}\left[Z^4\right] E^{3/4}\left[I^{4/3}_{[Z \ge 0]}\right] = \xi p^{3/4}.$$

Combining these three inequalities with (9.19) gives $s^2 \le p^{1/2}\xi^2 + (\xi p^{3/4})^{2/3}\xi^{4/3} = 2p^{1/2}\xi^2$. ∎

[†]For a related result, see Problem 25.19.

Chernoff's Theorem[†]

Theorem 9.3. *Let $X_1, X_2, \ldots$ be independent, identically distributed simple random variables satisfying $E[X_n] < 0$ and $P[X_n > 0] > 0$, let $M(t)$ be their common moment generating function, and put $\rho = \inf_t M(t)$. Then*

$$(9.20) \qquad \lim_{n \to \infty} \frac{1}{n} \log P[X_1 + \cdots + X_n \geq 0] = \log \rho.$$

PROOF. Put $Y_n = X_1 + \cdots + X_n$. Then $E[Y_n] < 0$ and $P[Y_n > 0] \geq P^n[X_1 > 0] > 0$, and so the hypotheses of Theorem 9 1 are satisfied. Define ρ_n and τ_n by $\inf_t M_n(t) = M_n(\tau_n) = \rho_n$, where $M_n(t)$ is the moment generating function of Y_n. Since $M_n(t) = M^n(t)$, it follows that $\rho_n = \rho^n$ and $\tau_n = \tau$, where $M(\tau) = \rho$.

Let Z_n be the analogue for Y_n of the Z described by (9.13). Its moment generating function (see (9.14)) is $M_n(\tau + t)/\rho^n = (M(\tau + t)/\rho)^n$. This is also the moment generating function of $V_1 + \cdots + V_n$ for independent random variables $V_1, \ldots, V_n$ each having moment generating function $M(\tau + t)/\rho$. Now each V_i has (see (9.15)) mean 0 and some positive variance σ^2 and fourth moment ξ^4 independent of i. Since Z_n must have the same moments as $V_1 + \cdots + V_n$, it has mean 0, variance $s_n^2 = n\sigma^2$, and fourth moment $\xi_n^4 = n\xi^4 + 3n(n-1)\sigma^4 = O(n^2)$ (see (6.2)). By Theorem 9.2, $P[Z_n \geq 0] \geq s_n^4/4\xi_n^4 \geq \alpha$ for some positive α independent of n. By Theorem 9.1 then, $P[Y_n \geq 0] = \rho^n e^{-\theta_n}$, where $0 \leq \theta_n \leq \tau_n s_n \alpha^{-1} - \log \alpha = \tau \alpha^{-1} \sigma \sqrt{n} - \log \alpha$. This gives (9.20), and shows, in fact, that the rate of convergence is $O(n^{-1/2})$. ∎

This result is important in the theory of statistical hypothesis testing. An informal treatment of the Bernoulli case will illustrate the connection.

Suppose $S_n = X_1 + \cdots + X_n$, where the X_i are independent and assume the values 1 and 0 with probabilities p and q. Now $P[S_n \geq na] = P[\sum_{k=1}^n (X_k - a) \geq 0]$, and Chernoff's theorem applies if $p < a < 1$. In this case $M(t) = E[e^{t(X_1 - a)}] = e^{-ta}(pe^t + q)$. Minimizing this shows that the ρ of Chernoff's theorem satisfies

$$-\log \rho = K(a, p) = a \log \frac{a}{p} + b \log \frac{b}{q},$$

where $b = 1 - a$. By (9.20), $n^{-1} \log P[S_n \geq na] \to -K(a, p)$; express this as

$$(9.21) \qquad P[S_n \geq na] \approx e^{-nK(a, p)}.$$

Suppose now that p is unknown and that there are two competing hypotheses concerning its value, the hypothesis H_1 that $p = p_1$ and the hypothesis H_2 that

---

[†]This theorem is not needed for the law of the iterated logarithm, Theorem 9.5.

$p = p_2$, where $p_1 < p_2$. Given the observed results $X_1, \ldots, X_n$ of n Bernoulli trials, one decides in favor of H_2 if $S_n \geq na$ and in favor of H_1 if $S_n < na$, where a is some number satisfying $p_1 < a < p_2$. The problem is to find an advantageous value for the threshold a.

By (9.21),

$$(9.22) \qquad P[S_n \geq na | H_1] \approx e^{-nK(a, p_1)},$$

where the notation indicates that the probability is calculated for $p = p_1$—that is, under the assumption of H_1. By symmetry,

$$(9\,23) \qquad P[S_n < na | H_2] = e^{-nK(a, p_2)}.$$

The left sides of (9.22) and (9 23) are the probabilities of erroneously deciding in favor of H_2 when H_1 is, in fact, true and of erroneously deciding in favor of H_1 when H_2 is, in fact, true—the probabilities describing the level and power of the test.

Suppose a is chosen so that $K(a, p_1) = K(a, p_2)$, which makes the two error probabilities approximately equal. This constraint gives for a a linear equation with solution

$$(9.24) \qquad a = a(p_1, p_2) = \frac{\log(q_1/q_2)}{\log(p_2/p_1) + \log(q_1/q_2)},$$

where $q_i = 1 - p_i$ The common error probability is approximately $e^{-nK(a, p_1)}$ for this value of a, and so the larger $K(a, p_1)$ is, the easier it is to distinguish statistically between p_1 and p_2.

Although $K(a(p_1, p_2), p_1)$ is a complicated function, it has a simple approximation for p_1 near p_2. As $x \to 0$, $\log(1 + x) = x - \frac{1}{2}x^2 + O(x^3)$. Using this in the definition of K and collecting terms gives

$$(9.25) \qquad K(p + x, p) = \frac{x^2}{2pq} + O(x^3), \qquad x \to 0.$$

Fix $p_1 = p$, and let $p_2 = p + t$; (9.24) becomes a function $\psi(t)$ of t, and expanding the logarithms gives

$$(9.26) \qquad \psi(t) = p + \tfrac{1}{2}t + O(t^2), \qquad t \to 0,$$

after some reductions. Finally, (9.25) and (9.26) together imply that

$$(9.27) \qquad K(\psi(t), p) = \frac{t^2}{8pq} + O(t^3), \qquad t \to 0.$$

In distinguishing $p_1 = p$ from $p_2 = p + t$ for small t, if a is chosen to equalize the two error probabilities, then their common value is about $e^{-nt^2/8pq}$. For t fixed, the nearer p is to $\frac{1}{2}$, the larger this probability is and the more difficult it is to distinguish p from $p + t$. As an example, compare $p = .1$ with $p = .5$. Now $.36nt^2/8(.1)(.9) = nt^2/8(.5)(.5)$. With a sample only 36 percent as large, .1 can therefore be distinguished from $.1 + t$ with about the same precision as .5 can be distinguished from $.5 + t$.

The Law of the Iterated Logarithm

The analysis of the rate at which S_n/n approaches the mean depends on the following variant of the theorem on large deviations.

Theorem 9.4.   *Let* $S_n = X_1 + \cdots + X_n$, *where the* X_n *are independent and identically distributed simple random variables with mean* 0 *and variance* 1. *If* a_n *are constants satisfying*

$$(9.28) \qquad\qquad a_n \to \infty, \qquad \frac{a_n}{\sqrt{n}} \to 0,$$

then

$$(9.29) \qquad\qquad P\left[S_n \geq a_n\sqrt{n} \right] = e^{-a_n^2(1+\zeta_n)/2}$$

for a sequence ζ_n *going to* 0.

PROOF.   Put $Y_n = S_n - a_n\sqrt{n} = \sum_{k=1}^{n}(X_k - a_n/\sqrt{n})$. Then $E[Y_n] < 0$. Since X_1 has mean 0 and variance 1, $P[X_1 > 0] > 0$, and it follows by (9.28) that $P[X_1 > a_n/\sqrt{n}] > 0$ for n sufficiently large, in which case $P[Y_n > 0] \geq P^n[X_1 - a_n/\sqrt{n} > 0] > 0$. Thus Theorem 9.1 applies to Y_n for all large enough n.

Let $M_n(t), \rho_n, \tau_n$, and Z_n be associated with Y_n as in the theorem. If $m(t)$ and $c(t)$ are the moment and cumulant generating functions of the X_n, then $M_n(t)$ is the nth power of the moment generating function $e^{-ta_n/\sqrt{n}}m(t)$ of $X_1 - a_n/\sqrt{n}$, and so Y_n has cumulant generating function

$$(9.30) \qquad\qquad C_n(t) = -ta_n\sqrt{n} + nc(t).$$

Since τ_n is the unique minimum of $C_n(t)$, and since $C_n'(t) = -a_n\sqrt{n} + nc'(t)$, τ_n is determined by the equation $c'(\tau_n) = a_n/\sqrt{n}$. Since X_1 has mean 0 and variance 1, it follows by (9.6) that

$$(9.31) \qquad\qquad c(0) = c'(0) = 0, \qquad c''(0) = 1.$$

Now $c'(t)$ is nondecreasing because $c(t)$ is convex, and since $c'(\tau_n) = a_n/\sqrt{n}$ goes to 0, τ_n must therefore go to 0 as well and must in fact be $O(a_n/\sqrt{n})$. By the second-order mean-value theorem for $c'(t), a_n/\sqrt{n} = c'(\tau_n) = \tau_n + O(\tau_n^2)$, from which follows

$$(9.32) \qquad\qquad \tau_n = \frac{a_n}{\sqrt{n}} + O\left(\frac{a_n^2}{n}\right).$$

By the third-order mean-value theorem for $c(t)$,

$$\log \rho_n = C_n(\tau_n) = -\tau_n a_n \sqrt{n} + nc(\tau_n)$$

$$= -\tau_n a_n \sqrt{n} + n\left[\tfrac{1}{2}\tau_n^2 + O(\tau_n^3)\right].$$

Applying (9.32) gives

$$(9.33) \qquad\qquad \log \rho_n = -\tfrac{1}{2}a_n^2 + o(a_n^2).$$

Now (see (9.14)) Z_n has moment generating function $M_n(\tau_n + t)/\rho_n$ and (see (9.30)) cumulant generating function $D_n(t) = C_n(\tau_n + t) - \log \rho_n = -(\tau_n + t)a_n\sqrt{n} + nc(t + \tau_n) - \log \rho_n$. The mean of Z_n is $D_n'(0) = 0$. Its variance s_n^2 is $D_n''(0)$; by (9.31), this is

$$(9.34) \qquad s_n^2 = nc''(\tau_n) = n(c''(0) + O(\tau_n)) = n(1 + o(1)).$$

The fourth cumulant of Z_n is $D_n''''(0) = nc''''(\tau_n) = O(n)$. By the formula (9.9) relating moments and cumulants (applicable because $E[Z_n] = 0$), $E[Z_n^4] = 3s_n^4 + D_n''''(0)$. Therefore, $E[Z_n^4]/s_n^4 \to 3$, and it follows by Theorem 9.2 that there exists an α such that $P[Z_n \ge 0] \ge \alpha > 0$ for all sufficiently large n.

By Theorem 9.1, $P[Y_n \ge 0] = \rho_n e^{-\theta_n}$ with $0 \le \theta_n \le \tau_n s_n \alpha^{-1} + \log \alpha$. By (9.28), (9.32), and (9.34), $\theta_n = O(a_n) = o(a_n^2)$, and it follows by (9.33) that $P[Y_n \ge 0] = e^{-a_n^2(1 + o(1))/2}$. ∎

The law of the iterated logarithm is this:

Theorem 9.5. *Let $S_n = X_1 + \cdots + X_n$, where the X_n are independent, identically distributed simple random variables with mean 0 and variance 1. Then*

$$(9.35) \qquad P\left[\limsup_n \frac{S_n}{\sqrt{2n\log\log n}} = 1\right] = 1.$$

Equivalent to (9.35) is the assertion that for positive ϵ

$$(9.36) \qquad P\left[S_n \ge (1 + \epsilon)\sqrt{2n\log\log n} \text{ i.o.}\right] = 0$$

and

$$(9.37) \qquad P\left[S_n \ge (1 - \epsilon)\sqrt{2n\log\log n} \text{ i.o.}\right] = 1.$$

The set in (9.35) is, in fact, the intersection over positive rational ϵ of the sets in (9.37) minus the union over positive rational ϵ of the sets in (9.36).

The idea of the proof is this. Write

$$(9.38) \qquad\qquad \phi(n) = \sqrt{2n \log\log n}\,.$$

If $A_n^{\pm} = [S_n \geq (1 \pm \epsilon)\phi(n)]$, then by (9.29), $P(A_n^{\pm})$ is near $(\log n)^{-(1 \pm \epsilon)^2}$. If n_k increases exponentially, say $n_k \sim \theta^k$ for $\theta > 1$, then $P(A_{n_k}^{\pm})$ is of the order $k^{-(1 \pm \epsilon)^2}$. Now $\Sigma_k k^{-(1 \pm \epsilon)^2}$ converges if the sign is $+$ and diverges if the sign is $-$. It will follow by the first Borel–Cantelli lemma that there is probability 0 that $A_{n_k}^{+}$ occurs for infinitely many k. In providing (9.36), an extra argument is required to get around the fact that the A_n^{+} for $n \neq n_k$ must also be accounted for (this requires choosing θ near 1). If the A_n^{-} were independent, it would follow by the second Borel–Cantelli lemma that with probability 1, $A_{n_k}^{-}$ occurs for infinitely many k, which would in turn imply (9.37). An extra argument is required to get around the fact that the $A_{n_k}^{-}$ are dependent (this requires choosing θ large).

For the proof of (9.36) a preliminary result is needed. Put $M_k = \max\{S_0, S_1, \ldots, S_k\}$, where $S_0 = 0$.

Theorem 9.6.   *If the X_k are independent simple random variables with mean 0 and variance 1, then for $\alpha \geq \sqrt{2}$.*

$$(9.39) \qquad\qquad P\!\left[\frac{M_n}{\sqrt{n}} \geq \alpha\right] \leq 2P\!\left[\frac{S_n}{\sqrt{n}} \geq \alpha - \sqrt{2}\,\right].$$

PROOF.   If $A_j = [M_{j-1} < \alpha\sqrt{n} \leq M_j]$, then

$$P\!\left[\frac{M_n}{\sqrt{n}} \geq \alpha\right] \leq P\!\left[\frac{S_n}{\sqrt{n}} \geq \alpha - \sqrt{2}\,\right] + \sum_{j=1}^{n-1} P\!\left(A_j \cap \left[\frac{S_n}{\sqrt{n}} \leq \alpha - \sqrt{2}\,\right]\right).$$

Since $S_n - S_j$ has variance $n - j$, it follows by independence and Chebyshev's inequality that the probability in the sum is at most

$$P\!\left(A_j \cap \left[\frac{|S_n - S_j|}{\sqrt{n}} > \sqrt{2}\,\right]\right) = P(A_j)P\!\left[\frac{|S_n - S_j|}{\sqrt{n}} > \sqrt{2}\,\right]$$

$$\leq P(A_j)\frac{n-j}{2n} \leq \tfrac{1}{2}P(A_j).$$

Since $\bigcup_{j=1}^{n-1} A_j \subset [M_n \geq \alpha\sqrt{n}\,]$,

$$P\!\left[\frac{M_n}{\sqrt{n}} \geq \alpha\right] \leq P\!\left[\frac{S_n}{\sqrt{n}} \geq \alpha - \sqrt{2}\,\right] + \tfrac{1}{2}P\!\left[\frac{M_n}{\sqrt{n}} \geq \alpha\right]. \qquad\blacksquare$$

PROOF OF (9.36). Given ϵ, choose θ so that $\theta > 1$ but $\theta^2 < 1 + \epsilon$. Let $n_k = \lfloor \theta^k \rfloor$ and $x_k = \theta(2\log\log n_k)^{1/2}$. By (9.29) and (9.39),

$$P\left[\frac{M_{n_k}}{\sqrt{n_k}} \geq x_k\right] \leq 2\exp\left[-\tfrac{1}{2}\left(x_k - \sqrt{2}\right)^2(1 + \xi_k)\right].$$

where $\xi_k \to 0$. The negative of the exponent is asymptotically $\theta^2 \log k$ and hence for large k exceeds $\theta \log k$, so that

$$P\left[\frac{M_{n_k}}{\sqrt{n_k}} \geq x_k\right] \leq \frac{2}{k^\theta}.$$

Since $\theta > 1$, it follows by the first Borel–Cantelli lemma that there is probability 0 that (see (9.38))

$$(9.40) \qquad\qquad M_{n_k} \geq \theta\phi(n_k)$$

for infinitely many k. Suppose that $n_{k-1} < n \leq n_k$ and that

$$(9.41) \qquad\qquad S_n > (1 + \epsilon)\phi(n).$$

Now $\phi(n) \geq \phi(n_{k-1}) \sim \theta^{-1/2}\phi(n_k)$; hence, by the choice of θ, $(1 + \epsilon)\phi(n) > \theta\phi(n_k)$ if k is large enough. Thus for sufficiently large k, (9.41) implies (9.40) (if $n_{k-1} < n \leq n_k$), and there is therefore probability 0 that (9.41) holds for infinitely many n.   ∎

PROOF OF (9.37). Given ϵ, choose an integer θ so large that $3\theta^{-1/2} < \epsilon$. Take $n_k = \theta^k$. Now $n_k - n_{k-1} \to \infty$, and (9.29) applies with $n = n_k - n_{k-1}$ and $a_n = x_k / \sqrt{n_k - n_{k-1}}$, where $x_k = (1 - \theta^{-1})\phi(n_k)$. It follows that

$$P\left[S_{n_k} - S_{n_{k-1}} \geq x_k\right] = P\left[S_{n_k - n_{k-1}} \geq x_k\right] = \exp\left[-\frac{1}{2}\frac{x_k^2}{n_k - n_{k-1}}(1 + \xi_k)\right],$$

where $\xi_k \to 0$. The negative of the exponent is asymptotically $(1 - \theta^{-1})\log k$ and so for large k is less than $\log k$, in which case $P[S_{n_k} - S_{n_{k-1}} \geq x_k] \geq k^{-1}$. The events here being independent, it follows by the second Borel–Cantelli lemma that with probability 1, $S_{n_k} - S_{n_{k-1}} \geq x_k$ for infinitely many k. On the other hand, by (9.36) applied to $\{-X_n\}$, there is probability 1 that $-S_{n_{k-1}} \leq 2\phi(n_{k-1}) \leq 2\theta^{-1/2}\phi(n_k)$ for all but finitely many k. These two inequalities give $S_{n_k} \geq x_k - 2\theta^{-1/2}\phi(n_k) > (1 - \epsilon)\phi(n_k)$, the last inequality because of the choice of θ.   ∎

That completes the proof of Theorem 9.5.

PROBLEMS

9.1. Prove (6.2) by using (9.9) and the fact that cumulants add in the presence of independence.

9.2. In the Bernoulli case, (9.21) gives

$$P[S_n \geq np + x_n] = \exp\left[-nK\left(p + \frac{x_n}{n}, p \right)(1 + o(1)) \right],$$

where $p < a < 1$ and $x_n = n(a - p)$. Theorem 9.4 gives

$$P[S_n \geq np + x_n] = \exp\left[-\frac{x_n^2}{2npq}(1 + o(1)) \right],$$

where $x_n = a_n\sqrt{npq}$. Resolve the apparent discrepancy. Use (9.25) to compare the two expressions in case x_n/n is small. See Problem 27.17.

9.3. Relabel the binomial parameter p as $\theta = f(p)$, where f is increasing and continuously differentiable. Show by (9.27) that the distinguishability of θ from $\theta + \Delta\theta$, as measured by K, is $(\Delta\theta)^2/8p(1 - p)(f'(p))^2 + O(\Delta\theta)^3$. The leading coefficient is independent of θ if $f(p) = \arcsin\sqrt{p}$.

9.4. From (9.35) and the same result for $\{-X_n\}$, together with the uniform boundedness of the X_n, deduce that with probability 1 the set of limit points of the sequence $\{S_n(2n \log\log n)^{-1/2}\}$ is the closed interval from -1 to $+1$.

9.5. ↑   Suppose X_n takes the values ± 1 with probability $\frac{1}{2}$ each, and show that $P[S_n = 0 \text{ i.o.}] = 1$. (This gives still another proof of the persistence of symmetric random walk on the line (Example 8.6).) Show more generally that, if the X_n are bounded by M, then $P[|S_n| \leq M \text{ i.o.}] = 1$.

9.6. Weakened versions of (9.36) are quite easy to prove. By a fourth-moment argument (see (6.2)), show that $P[S_n > n^{3/4}(\log n)^{(1+\epsilon)/4} \text{i.o.}] = 0$. Use (9.29) to give a simple proof that $P[S_n > (3n \log n)^{1/2} \text{i.o.}] = 0$.

9.7. Show that (9.35) is true if S_n is replaced by $|S_n|$ or $\max_{k \leq n} S_k$ or $\max_{k \leq n}|S_k|$.

CHAPTER 2

Measure

SECTION 10. GENERAL MEASURES

Lebesgue measure on the unit interval was central to the ideas in Chapter 1. Lebesgue measure on the entire real line is important in probability as well as in analysis generally, and a uniform treatment of this and other examples requires a notion of measure for which infinite values are possible. The present chapter extends the ideas of Sections 2 and 3 to this more general setting.

Classes of Sets

The σ-field of Borel sets in $(0, 1]$ played an essential role in Chapter 1, and it is necessary to construct the analogous classes for the entire real line and for k-dimensional Euclidean space.

Example 10.1. Let $x = (x_1, \ldots, x_k)$ be the generic point of Euclidean k-space R^k. The bounded rectangles

$$(10.1) \qquad \left[x = (x_1, \ldots, x_k) \colon a_i < x_i \le b_i, \, i = 1, \ldots, k \right]$$

will play in R^k the role intervals $(a, b]$ played in $(0, 1]$. Let $\mathscr{R}^k$ be the σ-field generated by these rectangles. This is the analogue of the class $\mathscr{B}$ of Borel sets in $(0, 1]$; see Example 2.6. The elements of $\mathscr{R}^k$ are the k-*dimensional Borel sets*. For $k = 1$ they are also called the *linear Borel sets*.

Call the rectangle (10.1) *rational* if the a_i and b_i are all rational. If G is an open set in R^k and $y \in G$, then there is a rational rectangle A_y such that $y \in A_y \subset G$. But then $G = \bigcup_{y \in G} A_y$, and since there are only countably many rational rectangles, this is a countable union. Thus $\mathscr{R}^k$ *contains the open sets*. Since a closed set has open complement, $\mathscr{R}^k$ also contains the *closed* sets. Just as $\mathscr{B}$ contains all the sets in $(0, 1]$ that actually arise in

158

ordinary analysis and probability theory, $\mathscr{R}^k$ contains all the sets in R^k that actually arise.

The σ-field $\mathscr{R}^k$ is generated by subclasses other than the class of rectangles. If A_n is the x-set where $a_i < x_i < b_i + n^{-1}$, $i = 1, \ldots, k$, then A_n is open and (10.1) is $\bigcap_n A_n$. Thus $\mathscr{R}^k$ *is generated by the open sets.* Similarly, it is generated by the closed sets. Now an open set is a countable union of rational rectangles. *Therefore, the (countable) class of rational rectangles generates $\mathscr{R}^k$.*   ∎

The σ-field $\mathscr{R}^1$ on the line R^1 is by definition generated by the finite intervals. The σ-field $\mathscr{B}$ in $(0,1]$ is generated by the subintervals of $(0,1]$. The question naturally arises whether the elements of $\mathscr{B}$ are the elements of $\mathscr{R}^1$ that happen to lie inside $(0,1]$, and the answer is yes. If $\mathscr{A}$ is a class of sets in a space Ω and Ω_0 is a subset of Ω, let $\mathscr{A} \cap \Omega_0 = [A \cap \Omega_0 \colon A \in \mathscr{A}]$.

Theorem 10.1.   (i)   *If $\mathscr{F}$ is a σ-field in Ω, then $\mathscr{F} \cap \Omega_0$ is a σ-field in Ω_0.*

(ii)   *If $\mathscr{A}$ generates the σ-field $\mathscr{F}$ in Ω, then $\mathscr{A} \cap \Omega_0$ generates the σ-field $\mathscr{F} \cap \Omega_0$ in Ω_0: $\sigma(\mathscr{A} \cap \Omega_0) = \sigma(\mathscr{A}) \cap \Omega_0$.*

PROOF.   Of course $\Omega_0 = \Omega \cap \Omega_0$ lies in $\mathscr{F} \cap \Omega_0$. If B lies in $\mathscr{F} \cap \Omega_0$, so that $B = A \cap \Omega_0$ for an $A \in \mathscr{F}$, then $\Omega_0 - B = (\Omega - A) \cap \Omega_0$ lies in $\mathscr{F} \cap \Omega_0$. If B_n lies in $\mathscr{F} \cap \Omega_0$ for all n, so that $B_n = A_n \cap \Omega_0$ for an $A_n \in \mathscr{F}$, then $\bigcup_n B_n = (\bigcup_n A_n) \cap \Omega_0$ lies in $\mathscr{F} \cap \Omega_0$. Hence part (i).

Let $\mathscr{F}_0$ be the σ-field $\mathscr{A} \cap \Omega_0$ generates in Ω_0. Since $\mathscr{A} \cap \Omega_0 \subset \mathscr{F} \cap \Omega_0$ and $\mathscr{F} \cap \Omega_0$ is a σ-field by part (i), $\mathscr{F}_0 \subset \mathscr{F} \cap \Omega_0$.

Now $\mathscr{F} \cap \Omega_0 \subset \mathscr{F}_0$ will follow if it is shown that $A \in \mathscr{F}$ implies $A \cap \Omega_0 \in \mathscr{F}_0$, or, to put it another way, if it is shown that $\mathscr{F}$ is contained in $\mathscr{G} = [A \subset \Omega \colon A \cap \Omega_0 \in \mathscr{F}_0]$. Since $A \in \mathscr{A}$ implies that $A \cap \Omega_0$ lies in $\mathscr{A} \cap \Omega_0$ and hence in $\mathscr{F} \cap \Omega_0$, it follows that $\mathscr{A} \subset \mathscr{G}$. It is therefore enough to show that $\mathscr{G}$ is a σ-field in Ω. Since $\Omega \cap \Omega_0 = \Omega_0$ lies in $\mathscr{F}_0$, it follows that $\Omega \in \mathscr{G}$. If $A \in \mathscr{G}$, then $(\Omega - A) \cap \Omega_0 = \Omega_0 - (A \cap \Omega_0)$ lies in $\mathscr{F}_0$ and hence $\Omega - A \in \mathscr{G}$. If $A_n \in \mathscr{G}$ for all n, then $(\bigcup_n A_n) \cap \Omega_0 = \bigcup_n (A_n \cap \Omega_0)$ lies in $\mathscr{F}_0$ and hence $\bigcup_n A_n \in \mathscr{G}$.   ∎

If $\Omega_0 \in \mathscr{F}$, then $\mathscr{F} \cap \Omega_0 = [A \colon A \subset \Omega_0, A \in \mathscr{F}]$. If $\Omega = R^1$, $\Omega_0 = (0,1]$, and $\mathscr{F} = \mathscr{R}^1$, and if $\mathscr{A}$ is the class of finite intervals on the line, then $\mathscr{A} \cap \Omega_0$ is the class of subintervals of $(0,1]$, and $\mathscr{B} = \sigma(\mathscr{A} \cap \Omega_0)$ is given by

$$(10.2) \qquad\qquad \mathscr{B} = \left[A \colon A \subset (0,1], A \in \mathscr{R}^1 \right].$$

A subset of $(0,1]$ is thus a Borel set (lies in $\mathscr{B}$) if and only if it is a linear Borel set (lies in $\mathscr{R}^1$), and the distinction in terminology can be dropped.

Conventions Involving ∞

Measures assume values in the set $[0,\infty]$ consisting of the ordinary nonnegative reals and the special value ∞, and some arithmetic conventions are called for.

For $x, y \in [0,\infty]$, $x \le y$ means that $y = \infty$ or else x and y are finite (that is, are ordinary real numbers) and $x \le y$ holds in the usual sense. Similarly, $x < y$ means that $y = \infty$ and x is finite or else x and y are both finite and $x < y$ holds in the usual sense.

For a finite or infinite sequence $x, x_1, x_2, \ldots$ in $[0,\infty]$,

$$(10.3) \qquad\qquad x = \sum_k x_k$$

means that either (i) $x = \infty$ and $x_k = \infty$ for some k, or (ii) $x = \infty$ and $x_k < \infty$ for all k and $\sum_k x_k$ is an ordinary divergent infinite series, or (iii) $x < \infty$ and $x_k < \infty$ for all k and (10.3) holds in the usual sense for $\sum_k x_k$ an ordinary finite sum or convergent infinite series. By these conventions and Dirichlet's theorem [A26], the order of summation in (10.3) has no effect on the sum.

For an infinite sequence $x, x_1, x_2, \ldots$ in $[0,\infty]$,

$$(10.4) \qquad\qquad x_k \uparrow x$$

means in the first place that $x_k \le x_{k+1} \le x$ and in the second place that either (i) $x < \infty$ and there is convergence in the usual sense, or (ii) $x_k = \infty$ for some k, or (iii) $x = \infty$ and the x_k are finite reals converging to infinity in the usual sense.

Measures

A set function μ on a field $\mathscr{F}$ in Ω is a *measure* if it satisfies these conditions:

(i) $\mu(A) \in [0,\infty]$ for $A \in \mathscr{F}$;

(ii) $\mu(\varnothing) = 0$;

(iii) if $A_1, A_2, \ldots$ is a disjoint sequence of $\mathscr{F}$-sets and if $\bigcup_{k=1}^{\infty} A_k \in \mathscr{F}$, then (see (10.3))

$$\mu\left(\bigcup_{k=1}^{\infty} A_k \right) = \sum_{k=1}^{\infty} \mu(A_k).$$

The measure μ is *finite* or *infinite* as $\mu(\Omega) < \infty$ or $\mu(\Omega) = \infty$; it is a *probability* measure if $\mu(\Omega) = 1$, as in Chapter 1.

If $\Omega = A_1 \cup A_2 \cup \ldots$ for some finite or countable sequence of $\mathscr{F}$-sets satisfying $\mu(A_k) < \infty$, then μ is *σ-finite*. The significance of this concept will be seen later. A finite measure is by definition σ-finite; a σ-finite measure may be finite or infinite. If $\mathscr{A}$ is a subclass of $\mathscr{F}$, then μ is *σ-finite on $\mathscr{A}$* if $\Omega = \bigcup_k A_k$ for some finite or infinite sequence of $\mathscr{A}$-sets satisfying $\mu(A_k) < \infty$. It is not required that these sets A_k be disjoint. Note that if Ω is not a finite or countable union of $\mathscr{A}$-sets, then *no* measure can be σ-finite on $\mathscr{A}$. It

is important to understand that σ-finiteness is a *joint* property of the space Ω, the measure μ, and the class $\mathscr{A}$.

If μ is a measure on a σ-field $\mathscr{F}$ in Ω, the triple $(\Omega, \mathscr{F}, \mu)$ is a *measure space*. (This term is not used if $\mathscr{F}$ is merely a field.) It is an infinite, a σ-finite, a finite, or a probability measure space according as μ has the corresponding property. If $\mu(A^c) = 0$ for an $\mathscr{F}$-set A, then A is a *support* of μ, and μ is *concentrated* on A. For a finite measure, A is a support if and only if $\mu(A) = \mu(\Omega)$.

The pair $(\Omega, \mathscr{F})$ itself is a *measurable space* if $\mathscr{F}$ is a σ-field in Ω. To say that μ is a measure on $(\Omega, \mathscr{F})$ indicates clearly both the space and the class of sets involved.

As in the case of probability measures, (iii) above is the condition of *countable additivity*, and it implies *finite additivity*: If $A_1, \ldots, A_n$ are disjoint $\mathscr{F}$-sets, then

$$\mu\left(\bigcup_{k=1}^{n} A_k\right) = \sum_{k=1}^{n} \mu(A_k).$$

As in the case of probability measures, if this holds for $n = 2$, then it extends inductively to all n.

Example 10.2.   A measure μ on $(\Omega, \mathscr{F})$ is *discrete* if there are finitely or countably many points ω_i in Ω and masses m_i in $[0, \infty]$ such that $\mu(A) = \sum_{\omega_i \in A} m_i$ for $A \in \mathscr{F}$. It is an infinite, a finite, or a probability measure as $\sum_i m_i$ diverges, or converges, or converges to 1; the last case was treated in Example 2.9. If $\mathscr{F}$ contains each singleton $\{\omega_i\}$, then μ is σ-finite if and only if $m_i < \infty$ for all i ■

Example 10.3.   Let $\mathscr{F}$ be the σ-field of all subsets of an arbitrary Ω, and let $\mu(A)$ be the number of points in A, where $\mu(A) = \infty$ if A is not finite. This μ is *counting measure*; it is finite if and only if Ω is finite, and is σ-finite if and only if Ω is countable. Even if $\mathscr{F}$ does not contain every subset of Ω, counting measure is well defined on $\mathscr{F}$.   ■

Example 10.4.   Specifying a measure includes specifying its domain. If μ is a measure on a field $\mathscr{F}$ and $\mathscr{F}_0$ is a field contained in $\mathscr{F}$, then the restriction μ_0 of μ to $\mathscr{F}_0$ is also a measure. Although often denoted by the same symbol, μ_0 is really a different measure from μ unless $\mathscr{F}_0 = \mathscr{F}$. Its properties may be different: If μ is counting measure on the σ-field $\mathscr{F}$ of all subsets of a countably infinite Ω, then μ is σ-finite, but its restriction to the σ-field $\mathscr{F}_0 = \{\varnothing, \Omega\}$ is not σ-finite.   ■

Certain properties of probability measures carry over immediately to the general case. First, μ is *monotone*: $\mu(A) \le \mu(B)$ if $A \subset B$. This is derived, just like its special case (2.5), from $\mu(A) + \mu(B - A) = \mu(B)$. But it is possible to go on and write $\mu(B - A) = \mu(B) - \mu(A)$ only if $\mu(B) < \infty$. If $\mu(B) = \infty$ and $\mu(A) < \infty$, then $\mu(B - A) = \infty$; but for every $\alpha \in [0, \infty]$ there are cases where $\mu(A) = \mu(B) = \infty$ and $\mu(B - A) = \alpha$. The inclusion-exclusion formula (2.9) also carries over without change to $\mathscr{F}$-sets of finite measure:

$$(10.5) \qquad \mu\left(\bigcup_{k=1}^{n} A_k\right) = \sum_{i} \mu(A_i) - \sum_{i<j} \mu(A_i \cap A_j) + \cdots$$

$$+ (-1)^{n+1} \mu(A_1 \cap \cdots \cap A_n).$$

The proof of *finite subadditivity* also goes through just as before:

$$\mu\left(\bigcup_{k=1}^{n} A_k\right) \le \sum_{k=1}^{n} \mu(A_k);$$

here the A_k need not have finite measure.

Theorem 10.2. *Let μ be a measure on a field $\mathscr{F}$.*

(i) *Continuity from below*: *If A_n and A lie in $\mathscr{F}$ and $A_n \uparrow A$, then*[†] $\mu(A_n) \uparrow \mu(A)$.

(ii) *Continuity from above*: *If A_n and A lie in $\mathscr{F}$ and $A_n \downarrow A$, and if $\mu(A_1) < \infty$, then $\mu(A_n) \downarrow \mu(A)$.*

(iii) *Countable subadditivity*: *If $A_1, A_2, \ldots$ and $\bigcup_{k=1}^{\infty} A_k$ lie in $\mathscr{F}$, then*

$$\mu\left(\bigcup_{k=1}^{\infty} A_k\right) \le \sum_{k=1}^{\infty} \mu(A_k).$$

(iv) *If μ is σ-finite on $\mathscr{F}$, then $\mathscr{F}$ cannot contain an uncountable, disjoint collection of sets of positive μ-measure.*

PROOF. The proofs of (i) and (iii) are exactly as for the corresponding parts of Theorem 2.1. The same is essentially true of (ii): If $\mu(A_1) < \infty$, subtraction is possible and $A_1 - A_n \uparrow A_1 - A$ implies that $\mu(A_1) - \mu(A_n) = \mu(A_1 - A_n) \uparrow \mu(A_1 - A) = \mu(A_1) - \mu(A)$.

There remains (iv). Let $[B_\theta : \theta \in \Theta]$ be a disjoint collection of $\mathscr{F}$-sets satisfying $\mu(B_\theta) > 0$. Consider an $\mathscr{F}$-set A for which $\mu(A) < \infty$. If $\theta_1, \ldots, \theta_n$ are distinct indices satisfying $\mu(A \cap B_{\theta_i}) \ge \epsilon > 0$, then $n\epsilon \le \sum_{i=1}^{n} \mu(A \cap B_{\theta_i}) \le \mu(A)$, and so $n \le \mu(A)/\epsilon$. Thus the index set $[\theta : \mu(A \cap B_\theta) > \epsilon]$ is finite,

[†]See (10.4).

and hence (take the union over positive rational ϵ) $[\theta: \mu(A \cap B_\theta) > 0]$ is countable. Since μ is σ-finite, $\Omega = \bigcup_k A_k$ for some finite or countable sequence of $\mathcal{F}$-sets A_k satisfying $\mu(A_k) < \infty$. But then $\Theta_k = [\theta: \mu(A_k \cap B_\theta) > 0]$ is countable for each k. Since $\mu(B_\theta) > 0$, there is a k for which $\mu(A_k \cap B_\theta) > 0$, and so $\Theta = \bigcup_k \Theta_k$: Θ is indeed countable. ∎

Uniqueness

According to Theorem 3.3, probability measures agreeing on a π-system $\mathcal{P}$ agree on $\sigma(\mathcal{P})$. There is an extension to the general case.

Theorem 10.3. *Suppose that μ_1 and μ_2 are measures on $\sigma(\mathcal{P})$, where $\mathcal{P}$ is a π-system, and suppose they are σ-finite on $\mathcal{P}$. If μ_1 and μ_2 agree on $\mathcal{P}$, then they agree on $\sigma(\mathcal{P})$.*

PROOF. Suppose that $B \in \mathcal{P}$ and $\mu_1(B) = \mu_2(B) < \infty$, and let $\mathcal{L}_B$ be the class of sets A in $\sigma(\mathcal{P})$ for which $\mu_1(B \cap A) = \mu_2(B \cap A)$. Then $\mathcal{L}_B$ is a λ-system containing $\mathcal{P}$ and hence (Theorem 3.2) containing $\sigma(\mathcal{P})$.

By σ-finiteness there exist $\mathcal{P}$-sets B_k satisfying $\Omega = \bigcup_k B_k$ and $\mu_1(B_k) = \mu_2(B_k) < \infty$. By the inclusion-exclusion formula (10.5),

$$\mu_\alpha \left(\bigcup_{i=1}^{n} (B_i \cap A) \right) = \sum_{1 \le i \le n} \mu_\alpha(B_i \cap A) - \sum_{1 \le i < j \le n} \mu_\alpha(B_i \cap B_j \cap A) + \cdots$$

for $\alpha = 1, 2$ and all n. Since $\mathcal{P}$ is a π-system containing the B_i, it contains the $B_i \cap B_j$, and so on. For each $\sigma(\mathcal{P})$-set A, the terms on the right above are therefore the same for $\alpha = 1$ as for $\alpha = 2$. The left side is then the same for $\alpha = 1$ as for $\alpha = 2$; letting $n \to \infty$ gives $\mu_1(A) = \mu_2(A)$. ∎

Theorem 10.4. *Suppose μ_1 and μ_2 are finite measures on $\sigma(\mathcal{P})$, where $\mathcal{P}$ is a π-system and Ω is a finite or countable union of sets in $\mathcal{P}$. If μ_1 and μ_2 agree on $\mathcal{P}$, then they agree on $\sigma(\mathcal{P})$.*

PROOF. By hypothesis, $\Omega = \bigcup_k B_k$ for $\mathcal{P}$-sets B_k, and of course $\mu_\alpha(B_k) \le \mu_\alpha(\Omega) < \infty$, $\alpha = 1, 2$. Thus μ_1 and μ_2 are σ-finite on $\mathcal{P}$, and Theorem 10.3 applies. ∎

Example 10.5. If $\mathcal{P}$ consists of the empty set alone, then it is a π-system and $\sigma(\mathcal{P}) = \{\varnothing, \Omega\}$. Any two finite measures agree on $\mathcal{P}$, but of course they need not agree on $\sigma(\mathcal{P})$. Theorem 10.4 does not apply in this case, because Ω is not a countable union of sets in $\mathcal{P}$. For the same reason, *no* measure on $\sigma(\mathcal{P})$ is σ-finite on $\mathcal{P}$, and hence Theorem 10.3 does not apply. ∎

Example 10.6. Suppose that $(\Omega, \mathscr{F}) = (R^1, \mathscr{R}^1)$ and $\mathscr{P}$ consists of the half-infinite intervals $(-\infty, x]$. By Theorem 10.4, two finite measures on $\mathscr{F}$ that agree on $\mathscr{P}$ also agree on $\mathscr{F}$. The $\mathscr{P}$-sets of finite measure required in the definition of σ-finiteness cannot in this example be made disjoint. ∎

Example 10.7. If a measure on $(\Omega, \mathscr{F})$ is σ-finite on a subfield $\mathscr{F}_0$ of $\mathscr{F}$, then $\Omega = \bigcup_k B_k$ for *disjoint* $\mathscr{F}_0$-sets B_k of finite measure: if they are not disjoint, replace B_k by $B_k \cap B_1^c \cdots \cap B_{k-1}^c$. ∎

The proof of Theorem 10.3 simplifies slightly if $\Omega = \bigcup_k B_k$ for disjoint $\mathscr{P}$-sets with $\mu_1(B_k) = \mu_2(B_k) < \infty$, because additivity itself can be used in place of the inclusion-exclusion formula.

PROBLEMS

10.1. Show that if conditions (i) and (iii) in the definition of measure hold, and if $\mu(A) < \infty$ for some $A \in \mathscr{F}$, then condition (ii) holds.

10.2. On the σ-field of all subsets of $\Omega = \{1, 2, \ldots\}$ put $\mu(A) = \Sigma_{k \in A} 2^{-k}$ if A is finite and $\mu(A) = \infty$ otherwise. Is μ finitely additive? Countably additive?

10.3. (a) In connection with Theorem 10.2(ii), show that if $A_n \downarrow A$ and $\mu(A_k) < \infty$ for some k, then $\mu(A_n) \downarrow \mu(A)$.

(b) Find an example in which $A_n \downarrow A$, $\mu(A_n) \equiv \infty$, and $A = \varnothing$.

10.4. The natural generalization of (4.9) is

(10.6) $$\mu\left(\liminf_n A_n\right) \le \liminf_n \mu(A_n)$$

$$\le \limsup_n \mu(A_n) \le \mu\left(\limsup_n A_n\right).$$

Show that the left-hand inequality always holds. Show that the right-hand inequality holds if $\mu(\bigcup_{k \ge n} A_k) < \infty$ for some n but can fail otherwise.

10.5. 3.10 A measure space $(\Omega, \mathscr{F}, \mu)$ is *complete* if $A \subset B$, $B \in \mathscr{F}$, and $\mu(B) = 0$ together imply that $A \in \mathscr{F}$—the definition is just as in the probability case. Use the ideas of Problem 3.10 to construct a complete measure space $(\Omega, \mathscr{F}^+, \mu^+)$ such that $\mathscr{F} \subset \mathscr{F}^+$ and μ and μ^+ agree on $\mathscr{F}$.

10.6. The condition in Theorem 10.2(iv) essentially characterizes σ-finiteness.

(a) Suppose that $(\Omega, \mathscr{F}, \mu)$ has no "infinite atoms," in the sense that for every A in $\mathscr{F}$, if $\mu(A) = \infty$, then there is in $\mathscr{F}$ a B such that $B \subset A$ and $0 < \mu(B) < \infty$. Show that if $\mathscr{F}$ does not contain an uncountable, disjoint collection of sets of positive measure, then μ is σ-finite. (Use Zorn's lemma.)

(b) Show by example that this is false without the condition that there are no "infinite atoms."

10.7. Example 10.5 shows that Theorem 10.3 fails without the σ-finiteness condition. Construct other examples of this kind.

SECTION 11. OUTER MEASURE

Outer Measure

An *outer measure* is a set function μ^* that is defined for all subsets of a space Ω and has these four properties:

(i) $\mu^*(A) \in [0, \infty]$ for every $A \subset \Omega$;

(ii) $\mu^*(\varnothing) = 0$;

(iii) μ^* is monotone: $A \subset B$ implies $\mu^*(A) \le \mu^*(B)$;

(iv) μ^* is countably subadditive: $\mu^*(\bigcup_n A_n) \le \sum_n \mu^*(A_n)$.

The set function P^* defined by (3.1) is an example, one which generalizes:

Example 11.1. Let ρ be a set function on a class $\mathscr{A}$ in Ω. Assume that $\varnothing \in \mathscr{A}$ and $\rho(\varnothing) = 0$, and that $\rho(A) \in [0, \infty]$ for $A \in \mathscr{A}$; ρ and $\mathscr{A}$ are otherwise arbitrary. Put

$$(11.1) \qquad \mu^*(A) = \inf \sum_n \rho(A_n),$$

where the infimum extends over all finite and countable coverings of A by $\mathscr{A}$-sets A_n. If no such covering exists, take $\mu^*(A) = \infty$ in accordance with the convention that the infimum over an empty set is ∞.

That μ^* satisfies (i), (ii), and (iii) is clear. If $\mu^*(A_n) = \infty$ for some n, then obviously $\mu^*(\bigcup_n A_n) \le \sum_n \mu^*(A_n)$. Otherwise, cover each A_n by $\mathscr{A}$-sets B_{nk} satisfying $\sum_k \rho(B_{nk}) < \mu^*(A_n) + \epsilon/2^n$; then $\mu^*(\bigcup_n A_n) \le \sum_{n,k} \rho(B_{nk}) < \sum_n \mu^*(A_n) + \epsilon$. Thus μ^* is an outer measure. ∎

Define A to be μ^*-*measurable* if

$$(11.2) \qquad \mu^*(A \cap E) + \mu^*(A^c \cap E) = \mu^*(E)$$

for every E. This is the general version of the definition (3.4) used in Section 3. By subadditivity it is equivalent to

$$(11.3) \qquad \mu^*(A \cap E) + \mu^*(A^c \cap E) \le \mu^*(E).$$

Denote by $\mathscr{M}(\mu^*)$ the class of μ^*-measurable sets.

The extension property for probability measures in Theorem 3.1 was proved by a sequence of lemmas the first three of which carry over directly to the case of the general outer measure: If P^* is replaced by μ^* and $\mathscr{M}$ by $\mathscr{M}(\mu^*)$ at each occurrence, the proofs hold word for word, symbol for symbol.

In particular, an examination of the arguments shows that ∞ as a possible value for μ^* does not require any changes. Lemma 3 in Section 3 becomes this:

Theorem 11.1. *If μ^* is an outer measure, then $\mathcal{M}(\mu^*)$ is a σ-field, and μ^* restricted to $\mathcal{M}(\mu^*)$ is a measure.*

This will be used to prove an extension theorem, but it has other applications as well.

Extension

Theorem 11.2. *A measure on a field has an extension to the generated σ-field.*

If the original measure on the field is σ-finite, then it follows by Theorem 10.3 that the extension is unique.

Theorem 11.2 can be deduced from Theorem 11.1 by the arguments used in the proof of Theorem 3.1.[†] It is unnecessary to retrace the steps, however, because the ideas will appear in stronger form in the proof of the next result, which generalizes Theorem 11.2.

Define a class $\mathcal{A}$ of subsets of Ω to be a *semiring* if

(i) $\varnothing \in \mathcal{A}$;

(ii) $A, B \in \mathcal{A}$ implies $A \cap B \in \mathcal{A}$;

(iii) if $A, B \in \mathcal{A}$ and $A \subset B$, then there exist disjoint $\mathcal{A}$-sets $C_1, \ldots, C_n$ such that $B - A = \bigcup_{k=1}^{n} C_k$.

The class of finite intervals in $\Omega = R^1$ and the class of subintervals of $\Omega = (0, 1]$ are the simplest examples of semirings. Note that a semiring need not contain Ω.

Theorem 11.3. *Suppose that μ is a set function on a semiring $\mathcal{A}$. Suppose that μ has values in $[0, \infty]$, that $\mu(\varnothing) = 0$, and that μ is finitely additive and countably subadditive. Then μ extends to a measure on $\sigma(\mathcal{A})$.*

This contains Theorem 11.2, because the conditions are all satisfied if $\mathcal{A}$ is a field and μ is a measure on it. If $\Omega = \bigcup_k A_k$ for a sequence of $\mathcal{A}$-sets satisfying $\mu(A_k) < \infty$, then it follows by Theorem 10.3 that the extension is unique.

[†]See also Problem 11.1.

PROOF. If A, B, and the C_k are related as in condition (iii) in the definition of semiring, then by finite additivity $\mu(B) = \mu(A) + \sum_{k=1}^{n} \mu(C_k) \geq \mu(A)$. Thus μ is monotone.

Define an outer measure μ^* by (11.1) for $\rho = \mu$:

$$(11.4) \qquad\qquad \mu^*(A) = \inf \sum_n \mu(A_n),$$

the infimum extending over coverings of A by $\mathscr{A}$-sets.

The first step is to show that $\mathscr{A} \subset \mathscr{M}(\mu^*)$. Suppose that $A \in \mathscr{A}$. If $\mu^*(E) = \infty$, then (11.3) holds trivially. If $\mu^*(E) < \infty$, for given ϵ choose $\mathscr{A}$-sets A_n such that $E \subset \bigcup_n A_n$ and $\sum_n \mu(A_n) < \mu^*(E) + \epsilon$. Since $\mathscr{A}$ is a semiring, $B_n = A \cap A_n$ lies in $\mathscr{A}$ and $A^c \cap A_n = A_n - B_n$ has the form $\bigcup_{k=1}^{m_n} C_{nk}$ for disjoint $\mathscr{A}$-sets C_{nk}. Note that $A_n = B_n \cup \bigcup_{k=1}^{m_n} C_{nk}$, where the union is disjoint, and that $A \cap E \subset \bigcup_n B_n$ and $A^c \cap E \subset \bigcup_n \bigcup_{k=1}^{m_n} C_{nk}$. By the definition of μ^* and the assumed finite additivity of μ,

$$\mu^*(A \cap E) + \mu^*(A^c \cap E) \leq \sum_n \mu(B_n) + \sum_n \sum_{k=1}^{m_n} \mu(C_{nk})$$

$$= \sum_n \mu(A_n) < \mu^*(E) + \epsilon.$$

Since ϵ is arbitrary, (11.3) follows. Thus $\mathscr{A} \subset \mathscr{M}(\mu^*)$.

The next step is to show that μ^* and μ agree on $\mathscr{A}$. If $A \subset \bigcup_n A_n$ for $\mathscr{A}$-sets A and A_n, then by the assumed countable subadditivity of μ and the monotonicity established above, $\mu(A) \leq \sum_n \mu(A \cap A_n) \leq \sum_n \mu(A_n)$. Therefore, $A \in \mathscr{A}$ implies that $\mu(A) \leq \mu^*(A)$ and hence, since the reverse inequality is an immediate consequence of (11.4), $\mu(A) = \mu^*(A)$. Thus μ^* agrees with μ on $\mathscr{A}$.

Since $\mathscr{A} \subset \mathscr{M}(\mu^*)$ and $\mathscr{M}(\mu^*)$ is a σ-field (Theorem 11.1),

$$\mathscr{A} \subset \sigma(\mathscr{A}) \subset \mathscr{M}(\mu^*) \subset 2^\Omega.$$

Since μ^* is countably additive when restricted to $\mathscr{M}(\mu^*)$ (Theorem 11.1 again), μ^* further restricted to $\sigma(\mathscr{A})$ is an extension of μ on $\mathscr{A}$, as required.   ∎

Example 11.2. For $\mathscr{A}$ take the semiring of subintervals of $\Omega = (0, 1]$ (together with the empty set). For μ take length λ: $\lambda(a, b] = b - a$. The finite additivity and countable subadditivity of λ follow by Theorem 1.3.[†] By Theorem 11.3, λ extends to a measure on the class $\sigma(\mathscr{A}) = \mathscr{B}$ of Borel sets in $(0, 1]$.   ∎

[†]On a field, countable additivity implies countable subadditivity, and λ is in fact countably additive on $\mathscr{A}$—but $\mathscr{A}$ is merely a semiring. Hence the separate consideration of additivity and subadditivity; but see Problem 11.2.

This gives a second construction of Lebesgue measure in the unit interval. In the first construction λ was extended first from the class of intervals to the field $\mathcal{B}_0$ of finite disjoint unions of intervals (see Theorem 2.2) and then by Theorem 11.2 (in its special form Theorem 3.1) from $\mathcal{B}_0$ to $\mathcal{B} = \sigma(\mathcal{B}_0)$. Using Theorem 11.3 instead of Theorem 11.2 effects a slight economy, since the extension then goes from $\mathcal{A}$ directly to $\mathcal{B}$ without the intermediate stop at $\mathcal{B}_0$, and the arguments involving (2.13) and (2.14) become unnecessary.

Example 11.3. In Theorem 11.3 take for $\mathcal{A}$ the semiring of finite intervals on the real line R^1, and consider $\lambda_1(a, b] = b - a$. The arguments for Theorem 1.3 in no way require that the (finite) intervals in question be contained in $(0, 1]$, and so λ_1 is finitely additive and countably subadditive on this class $\mathcal{A}$. Hence λ_1 extends to the σ-field $\mathscr{R}^1$ of linear Borel sets, which is by definition generated by $\mathcal{A}$. This defines Lebesgue measure λ_1 over the whole real line. ∎

A subset of $(0, 1]$ lies in $\mathcal{B}$ if and only if it lies in $\mathscr{R}^1$ (see (10.2)). Now $\lambda_1(A) = \lambda(A)$ for subintervals A of $(0.1]$, and it follows by uniqueness (Theorem 3.3) that $\lambda_1(A) = \lambda(A)$ for all A in $\mathcal{B}$. Thus there is no inconsistency in dropping λ_1 and using λ to denote Lebesgue measure on $\mathscr{R}^1$ as well as on $\mathcal{B}$.

Example 11.4. The class of bounded rectangles in R^k is a semiring, a fact needed in the next section. Suppose that $A = [x : x_i \in I_i, i \le k]$ and $B = [x : x_i \in J_i, i \le k]$ are nonempty rectangles, the I_i and J_i being finite intervals. If $A \subset B$, then $I_i \subset J_i$, so that $J_i - I_i$ is a disjoint union $I_i' \cup I_i''$ of intervals (possibly empty). Consider the 3^k disjoint rectangles $[x : x_i \in U_i, i \le k]$, where for each i, U_i is I_i or I_i' or I_i''. One of these rectangles is A itself, and $B - A$ is the union of the others. The rectangles thus form a semiring. ∎

An Approximation Theorem

If $\mathcal{A}$ is a semiring, then by Theorem 10.3 a measure on $\sigma(\mathcal{A})$ is determined by its values on $\mathcal{A}$ if it is σ-finite there. Theorem 11.4 shows more explicitly how the measure of a $\sigma(\mathcal{A})$-set can be approximated by the measures of $\mathcal{A}$-sets.

Lemma 1. *If $A, A_1, \ldots, A_n$ are sets in a semiring $\mathcal{A}$, then there are disjoint $\mathcal{A}$-sets $C_1, \ldots, C_m$ such that*

$$A \cap A_1^c \cap \cdots \cap A_n^c = C_1 \cup \cdots \cup C_m.$$

PROOF. The case $n = 1$ follows from the definition of semiring applied to $A \cap A_1^c = A - (A \cap A_1)$. If the result holds for n, then $A \cap A_1^c \cap \cdots \cap A_{n+1}^c = \bigcup_{j=1}^m (C_j \cap A_{n+1}^c)$; apply the case $n = 1$ to each set in the union. ∎

Theorem 11.4. *Suppose that $\mathscr{A}$ is a semiring, μ is a measure on $\mathscr{F} = \sigma(\mathscr{A})$, and μ is σ-finite on $\mathscr{A}$.*

(i) *If $B \in \mathscr{F}$ and $\epsilon > 0$, there exists a finite or infinite disjoint sequence $A_1, A_2, \ldots$ of $\mathscr{A}$-sets such that $B \subset \bigcup_k A_k$ and $\mu((\bigcup_k A_k) - B) < \epsilon$.*

(ii) *If $B \in \mathscr{F}$ and $\epsilon > 0$, and if $\mu(B) < \infty$, then there exists a finite disjoint sequence $A_1, \ldots, A_n$ of $\mathscr{A}$-sets such that $\mu(B \vartriangle (\bigcup_{k=1}^n A_k)) < \epsilon$.*

PROOF. Return to the proof of Theorem 11.3. If μ^* is the outer measure defined by (11.4), then $\mathscr{F} \subset \mathscr{M}(\mu^*)$ and μ^* agrees with μ on $\mathscr{A}$, as was shown. Since μ^* restricted to $\mathscr{F}$ is a measure, it follows by Theorem 10.3 that μ^* agrees with μ on $\mathscr{F}$ as well.

Suppose now that B lies in $\mathscr{F}$ and $\mu(B) = \mu^*(B) < \infty$. There exist $\mathscr{A}$-sets A_k such that $B \subset \bigcup_k A_k$ and $\mu(\bigcup_k A_k) \leq \Sigma_k \mu(A_k) < \mu(B) + \epsilon$; but then $\mu((\bigcup_k A_k) - B) < \epsilon$. To make the sequence $\{A_k\}$ disjoint, replace A_k by $A_k \cap A_1^c \cap \cdots \cap A_{k-1}^c$; by Lemma 1, each of these sets is a finite disjoint union of sets in $\mathscr{A}$.

Next suppose that B lies in $\mathscr{F}$ and $\mu(B) = \mu^*(B) = \infty$. By σ-finiteness there exist $\mathscr{A}$-sets C_m such that $\Omega = \bigcup_m C_m$ and $\mu(C_m) < \infty$. By what has just been shown, there exist $\mathscr{A}$-sets A_{mk} such that $B \cap C_m \subset \bigcup_k A_{mk}$ and $\mu((\bigcup_k A_{mk}) - (B \cap C_m)) < \epsilon/2^m$. The sets A_{mk} taken all together provide a sequence $A_1, A_2, \ldots$ of $\mathscr{A}$-sets satisfying $B \subset \bigcup_k A_k$ and $\mu((\bigcup_k A_k) - B) < \epsilon$. As before, the A_k can be made disjoint.

To prove part (ii), consider the A_k of part (i). If B has finite measure, so has $A = \bigcup_k A_k$, and hence by continuity from above (Theorem 10.2(ii)), $\mu(A - \bigcup_{k \leq n} A_k) < \epsilon$ for some n. But then $\mu(B \vartriangle (\bigcup_{k=1}^n A_k)) < 2\epsilon$. ∎

If, for example, B is a linear Borel set of finite Lebesgue measure, then $\lambda(B \vartriangle (\bigcup_{k=1}^n A_k)) < \epsilon$ for some disjoint collection of finite intervals $A_1, \ldots, A_n$.

Corollary 1. *If μ is a finite measure on a σ-field $\mathscr{F}$ generated by a field $\mathscr{F}_0$, then for each $\mathscr{F}$-set A and each positive ϵ there is an $\mathscr{F}_0$-set B such that $\mu(A \vartriangle B) < \epsilon$.*

PROOF. This is of course an immediate consequence of part (ii) of the theorem, but there is a simple direct argument. Let $\mathscr{G}$ be the class of $\mathscr{F}$-sets with the required property. Since $A^c \vartriangle B^c = A \vartriangle B$, $\mathscr{G}$ is closed under complementation. If $A = \bigcup_n A_n$, where $A_n \in \mathscr{G}$, given ϵ choose n_0 so that $\mu(A - \bigcup_{n \leq n_0} A_n) < \epsilon$, and then choose $\mathscr{F}_0$-sets B_n, $n \leq n_0$, so that $\mu(A_n \vartriangle B_n) < \epsilon/n_0$. Since $(\bigcup_{n \leq n_0} A_n) \vartriangle (\bigcup_{n \leq n_0} B_n) \subset \bigcup_{n \leq n_0} (A_n \vartriangle B_n)$, the $\mathscr{F}_0$-set $B = \bigcup_{n \leq n_0} B_n$ satisfies $\mu(A \vartriangle B) < 2\epsilon$. Of course $\mathscr{F}_0 \subset \mathscr{G}$; since $\mathscr{G}$ is a σ-field, $\mathscr{F} \subset \mathscr{G}$, as required. ∎

Corollary 2. *Suppose that $\mathscr{A}$ is a semiring, Ω is a countable union of $\mathscr{A}$-sets, and μ_1, μ_2 are measures on $\mathscr{F} = \sigma(\mathscr{A})$. If $\mu_1(A) \leq \mu_2(A) < \infty$ for $A \in \mathscr{A}$, then $\mu_1(B) \leq \mu_2(B)$ for $B \in \mathscr{F}$.*

PROOF. Since μ_2 is σ-finite on $\mathscr{A}$, the theorem applies. If $\mu_2(B) < \infty$, choose disjoint $\mathscr{A}$-sets A_k such that $B \subset \bigcup_k A_k$ and $\sum_k \mu_2(A_k) < \mu_2(B) + \epsilon$. Then $\mu_1(B) \leq \sum_k \mu_1(A_k) \leq \sum_k \mu_2(A_k) < \mu_2(B) + \epsilon$.   ∎

A fact used in the next section:

Lemma 2. *Suppose that* μ *is a nonnegative and finitely additive set function on a semiring* $\mathscr{A}$, *and let* $A, A_1, \ldots, A_n$ *be sets in* $\mathscr{A}$.

(i) *If* $\bigcup_{i=1}^n A_i \subset A$ *and the* A_i *are disjoint, then* $\sum_{i=1}^n \mu(A_i) \leq \mu(A)$.

(ii) *If* $A \subset \bigcup_{i=1}^n A_i$, *then* $\mu(A) \leq \sum_{i=1}^n \mu(A_i)$.

PROOF. For part (i), use Lemma 1 to choose disjoint $\mathscr{A}$-sets C_j such that $A - \bigcup_{i=1}^n A_i = \bigcup_{j=1}^m C_j$. Since μ is finitely additive and nonnegative, it follows that $\mu(A) = \sum_{i=1}^n \mu(A_i) + \sum_{j=1}^m \mu(C_j) \geq \sum_{i=1}^n \mu(A_i)$.

For (ii), take $B_1 = A \cap A_1$ and $B_i = A \cap A_i \cap A_1^c \cap \cdots \cap A_{i-1}^c$ for $i > 1$. By Lemma 1, each B_i is a finite disjoint union of $\mathscr{A}$-sets C_{ij}. Since the B_i are disjoint, $A = \bigcup_i B_i = \bigcup_{ij} C_{ij}$ and $\bigcup_j C_{ij} \subset A_i$, it follows by finite additivity and part (i) that $\mu(A) = \sum_i \sum_j \mu(C_{ij}) \leq \sum_i \mu(A_i)$.   ∎
Compare Theorem 1.3.

PROBLEMS

11.1. The proof of Theorem 3.1 obviously applies if the probability measure is replaced by a finite measure, since this is only a matter of rescaling. Take as a starting point then the fact that a finite measure on a field extends uniquely to the generated σ-field. By the following steps, prove Theorem 11.2—that is, remove the assumption of finiteness.

(a) Let μ be a measure (not necessarily even σ-finite) on a field $\mathscr{F}_0$, and let $\mathscr{F} = \sigma(\mathscr{F}_0)$. If A is a nonempty set in $\mathscr{F}_0$ and $\mu(A) < \infty$, restrict μ to a finite measure μ_A on the field $\mathscr{F}_0 \cap A$, and extend μ_A to a finite measure $\hat{\mu}_A$ on the σ-field $\mathscr{F} \cap A$ generated in A by $\mathscr{F}_0 \cap A$.

(b) Suppose that $E \in \mathscr{F}$. If there exist disjoint $\mathscr{F}_0$-sets A_n such that $E \subset \bigcup_n A_n$ and $\mu(A_n) < \infty$, put $\hat{\mu}(E) = \sum_n \hat{\mu}_{A_n}(E \cap A_n)$ and prove consistency. Otherwise put $\hat{\mu}(E) = \infty$.

(c) Show that $\hat{\mu}$ is a measure on $\mathscr{F}$ and agrees with μ on $\mathscr{F}_0$.

11.2. Suppose that μ is a nonnegative and finitely additive set function on a semiring $\mathscr{A}$.

(a) Use Lemmas 1 and 2, without reference to Theorem 11.3, to show that μ is countably subadditive if and only if it is countably additive.

(b) Find an example where μ is not countably subadditive.

11.3. Show that Theorem 11.4(ii) can fail if $\mu(B) = \infty$.

11.4. This and Problems 11.5, 16.12, 17.12, 17.13, and 17.14 lead to proofs of the *Daniell-Stone* and *Riesz representation theorems*.

Let Λ be a real linear functional on a vector lattice $\mathcal{L}$ of (finite) real functions on a space Ω. This means that if f and g lie in $\mathcal{L}$, then so do $f \vee g$ and $f \wedge g$ (with values $\max\{f(\omega), g(\omega)\}$ and $\min\{f(\omega), g(\omega)\}$), as well as $\alpha f + \beta g$, and $\Lambda(\alpha f + \beta g) = \alpha \Lambda(f) + \beta \Lambda(g)$. Assume further of $\mathcal{L}$ that $f \in \mathcal{L}$ implies $f \wedge 1 \in \mathcal{L}$ (where 1 denotes the function identically equal to 1). Assume further of Λ that it is positive in the sense that $f \geq 0$ (pointwise) implies $\Lambda(f) \geq 0$ and continuous from above at 0 in the sense that $f_n \downarrow 0$ (pointwise) implies $\Lambda(f_n) \to 0$.

(a) If $f \leq g$ ($f, g \in \mathcal{L}$), define in $\Omega \times R^1$ an "interval"

$$(11 5) \qquad (f, g] = \left[(\omega, t): f(\omega) < t \leq g(\omega) \right].$$

Show that these sets form a semiring $\mathcal{A}_0$.

(b) Define a set function ν_0 on $\mathcal{A}_0$ by

$$(11.6) \qquad \nu_0(f, g] = \Lambda(g - f).$$

Show that ν_0 is finitely additive and countably subadditive on $\mathcal{A}_0$.

11.5. ↑   (a) Assume $f \in \mathcal{L}$ and let $f_n = (n(f - f \wedge 1)) \wedge 1$. Show that $f(\omega) \leq 1$ implies $f_n(\omega) = 0$ for all n and $f(\omega) > 1$ implies $f_n(\omega) = 1$ for all sufficiently large n. Conclude that for $x > 0$,

$$(11.7) \qquad (0, xf_n] \uparrow \left[\omega. f(\omega) > 1 \right] \times (0, x].$$

(b) Let $\mathcal{F}$ be the smallest σ-field with respect to which every f in $\mathcal{L}$ is measurable: $\mathcal{F} = \sigma[f^{-1}H: f \in \mathcal{L}, H \in \mathcal{R}^1]$. Let $\mathcal{F}_0$ be the class of A in $\mathcal{F}$ for which $A \times (0, 1] \in \sigma(\mathcal{A}_0)$. Show that $\mathcal{F}_0$ is a semiring and that $\mathcal{F} = \sigma(\mathcal{F}_0)$.

(c) Let ν be the extension of ν_0 (see (11.6)) to $\sigma(\mathcal{A}_0)$, and for $A \in \mathcal{F}_0$ define $\mu_0(A) = \nu(A \times (0, 1])$. Show that μ_0 is finitely additive and countably subadditive on the semiring $\mathcal{F}_0$.

SECTION 12. MEASURES IN EUCLIDEAN SPACE

Lebesgue Measure

In Example 11.3 Lebesgue measure λ was constructed on the class $\mathcal{R}^1$ of linear Borel sets. By Theorem 10.3, λ is the only measure on $\mathcal{R}^1$ satisfying $\lambda(a, b] = b - a$ for all intervals. There is in k-space an analogous k-dimensional Lebesgue measure λ_k on the class $\mathcal{R}^k$ of k-dimensional Borel sets (Example 10.1). It is specified by the requirement that bounded rectangles have measure

$$(12.1) \qquad \lambda_k[x: a_i < x_i \leq b_i, i = 1, \ldots, k] = \prod_{i=1}^{k} (b_i - a_i).$$

This is ordinary volume—that is, length ($k = 1$), area ($k = 2$), volume ($k = 3$), or hypervolume ($k \geq 4$).

Since an intersection of rectangles is again a rectangle, the uniqueness theorem shows that (12.1) completely determines λ_k. That there does exist such a measure on $\mathscr{R}^k$ can be proved in several ways. One is to use the ideas involved in the case $k = 1$. A second construction is given in Theorem 12.5. A third, independent, construction uses the general theory of product measures; this is carried out in Section 18.[†] For the moment, assume the existence on $\mathscr{R}^k$ of a measure λ_k satisfying (12.1). Of course, λ_k is σ-finite.

A basic property of λ_k is *translation invariance*.[‡]

Theorem 12.1. *If $A \in \mathscr{R}^k$, then $A + x = [a + x: \ a \in A] \in \mathscr{R}^k$ and $\lambda_k(A) = \lambda_k(A + x)$ for all x.*

PROOF. If $\mathscr{G}$ is the class of A such that $A + x$ is in $\mathscr{R}^k$ for all x, then $\mathscr{G}$ is a σ-field containing the bounded rectangles, and so $\mathscr{G} \supset \mathscr{R}^k$. Thus $A + x \in \mathscr{R}^k$ for $A \in \mathscr{R}^k$.

For fixed x define a measure μ on $\mathscr{R}^k$ by $\mu(A) = \lambda_k(A + x)$. Then μ and λ_k agree on the π-system of bounded rectangles and so agree for all Borel sets.   ■

If A is a $(k - 1)$-dimensional subspace and x lies outside A, the hyperplanes $A + tx$ for real t are disjoint, and by Theorem 12.1, all have the same measure. Since only countably many disjoint sets can have positive measure (Theorem 10.2(iv)), the measure common to the $A + tx$ must be 0. *Every $(k - 1)$-dimensional hyperplane has k-dimensional Lebesgue measure* 0.

The Lebesgue measure of a rectangle is its ordinary volume. The following theorem makes it possible to calculate the measures of simple figures.

Theorem 12.2. *If $T: R^k \to R^k$ is linear and nonsingular, then $A \in \mathscr{R}^k$ implies that $TA \in \mathscr{R}^k$ and*

$$(12.2) \qquad\qquad \lambda_k(TA) = |\det T| \cdot \lambda_k(A).$$

Since a parallelepiped is the image of a rectangle under a linear transformation, (12.2) can be used to compute its volume. If T is a rotation or a reflection—an orthogonal or a unitary transformation—then $\det T = \pm 1$, and so $\lambda_k(TA) = \lambda_k(A)$. Hence every rigid transformation or isometry (an orthogonal transformation followed by a translation) preserves Lebesgue measure. An affine transformation has the form $Fx = Tx + x_0$ (the general

linear transformation T followed by a translation); it is nonsingular if T is. It follows by Theorems 12.1 and 12.2 that $\lambda_k(FA) = |\det T| \cdot \lambda_k(A)$ in the nonsingular case.

PROOF OF THE THEOREM. Since $T \bigcup_n A_n = \bigcup_n TA_n$ and $TA^c = (TA)^c$ because of the assumed nonsingularity of T, the class $\mathscr{G} = [A: TA \in \mathscr{R}^k]$ is a σ-field. Since TA is open for open A, it follows again by the assumed nonsingularity of T that $\mathscr{G}$ contains all the open sets and hence (Example 10.1) all the Borel sets. Therefore, $A \in \mathscr{R}^k$ implies $TA \in \mathscr{R}^k$.

For $A \in \mathscr{R}^k$, set $\mu_1(A) = \lambda_k(TA)$ and $\mu_2(A) = |\det T| \cdot \lambda_k(A)$. Then μ_1 and μ_2 are measures, and by Theorem 10.3 they will agree on $\mathscr{R}^k$ (which is the assertion (12.2)) if they agree on the π-system consisting of the rectangles $[x: a_i < x_i \le b_i, \ i = 1, \ldots, k]$ for which the a_i and the b_i are all rational (Example 10.1). It suffices therefore to prove (12.2) for rectangles with sides of rational length. Since such a rectangle is a finite disjoint union of cubes and λ_k is translation-invariant, it is enough to check (12.2) for cubes

$$(12.3) \qquad A = [x: 0 < x_i \le c, i = 1, \ldots, k]$$

that have their lower corner at the origin.

Now the general T can by elementary row and column operations[†] be represented as a product of linear transformations of these three special forms:

(1°) $T(x_1, \ldots, x_k) = (x_{\pi 1}, \ldots, x_{\pi k})$, where π is a permutation of the set $\{1, 2, \ldots, k\}$;

(2°) $T(x_1, \ldots, x_k) = (\alpha x_1, x_2, \ldots, x_k)$;

(3°) $T(x_1, \ldots, x_k) = (x_1 + x_2, x_2, \ldots, x_k)$.

Because of the rule for multiplying determinants, it suffices to check (12.2) for T of these three forms. And, as observed, for each such T it suffices to consider cubes (12.3).

(1°): Such a T is a permutation matrix, and so $\det T = \pm 1$. Since (12.3) is invariant under T, (12.2) is in this case obvious.

(2°): Here $\det T = \alpha$, and $TA = [x: x_1 \in H, \ 0 < x_i \le c, \ i = 2, \ldots, k]$, where $H = (0, \alpha c]$ if $\alpha > 0$, $H = \{0\}$ if $\alpha = 0$ (although α cannot in fact be 0 if T is nonsingular), and $H = [\alpha c, 0)$ if $\alpha < 0$. In each case, $\lambda_k(TA) = |\alpha| \cdot c^k = |\alpha| \cdot \lambda_k(A)$.

[†]BIRKHOFF & MAC LANE, Section 8.9
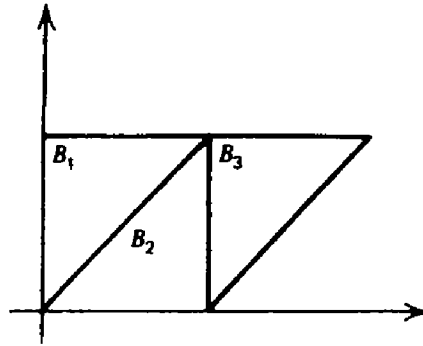
(3°): Here $\det T = 1$. Let $B = [x: 0 < x_i \le c, i = 3, \ldots, k]$, where $B = R^k$ if $k < 3$, and define

$$B_1 = [x: 0 < x_1 \le x_2 \le c] \cap B,$$
$$B_2 = [x: 0 < x_2 < x_1 \le c] \cap B,$$
$$B_3 = [x: c < x_1 \le c + x_2, 0 < x_2 \le c] \cap B.$$

Then $A = B_1 \cup B_2$, $TA = B_2 \cup B_3$, and $B_1 + (c, 0, \ldots, 0) = B_3$. Since $\lambda_k(B_1) = \lambda_k(B_3)$ by translation invariance, (12.2) follows by additivity.   ∎

If T is singular, then $\det T = 0$ and TA lies in a $(k-1)$-dimensional subspace. Since such a subspace has measure 0, (12.2) holds if A and TA lie in $\mathscr{R}^k$. The surprising thing is that $A \in \mathscr{R}^k$ need not imply that $TA \in \mathscr{R}^k$ if T is singular. Even for a very simple transformation such as the projection $T(x_1, x_2) = (x_1, 0)$ in the plane, there exist Borel sets A for which TA is not a Borel set.[†]

Regularity

Important among measures on $\mathscr{R}^k$ are those assigning finite measure to bounded sets. They share with λ_k the property of *regularity*:

Theorem 12.3. *Suppose that μ is a measure on $\mathscr{R}^k$ such that $\mu(A) < \infty$ if A is bounded.*

(i) *For $A \in \mathscr{R}^k$ and $\epsilon > 0$, there exist a closed C an open G such that $C \subset A \subset G$ and $\mu(G - C) < \epsilon$.*

(ii) *If $\mu(A) < \infty$, then $\mu(A) = \sup \mu(K)$, the supremum extending over the compact subsets K of A.*

Proof. The second part of the theorem follows form the first: $\mu(A) < \infty$ implies that $\mu(A - A_0) < \epsilon$ for a bounded subset A_0 of A, and it then follows from the first part that $\mu(A_0 - K) < \epsilon$ for a closed and hence compact subset K of A_0.

[†]See Hausdorff, p. 241.

To prove (i) consider first a bounded rectangle $A = [x: a_i < x_i \le b_i, i \le k]$. The set $G_n = [x: a_i < x_i < b_i + n^{-1}, i \le k]$ is open and $G_n \downarrow A$. Since $\mu(G_1)$ is finite by hypothesis, it follows by continuity from above that $\mu(G_n - A) < \epsilon$ for large n. A bounded rectangle can therefore be approximated from the outside by open sets.

The rectangles form a semiring (Example 11.4). For an arbitrary set A in $\mathscr{R}^k$, by Theorem 11.4(i) there exist bounded rectangles A_k such that $A \subset \bigcup_k A_k$ and $\mu((\bigcup_k A_k) - A) < \epsilon$. Choose open sets G_k such that $A_k \subset G_k$ and $\mu(G_k - A_k) < \epsilon/2^k$. Then $G = \bigcup_k G_k$ is open and $\mu(G - A) < 2\epsilon$. Thus the general k-dimensional Borel set can be approximated from the outside by open sets. To approximate from the inside by closed sets, pass to complements.   ∎

Specifying Measures on the Line

There are on the line many measures other than λ that are important for probability theory. There is a useful way to describe the collection of all measures on $\mathscr{R}^1$ that assign finite measure to each bounded set.

If μ is such a measure, define a real function F by

$$(12.4) \qquad F(x) = \begin{cases} \mu(0, x] & \text{if } x \ge 0, \\ -\mu(x, 0] & \text{if } x \le 0. \end{cases}$$

It is because $\mu(A) < \infty$ for bounded A that F is a finite function. Clearly, F is nondecreasing. Suppose that $x_n \downarrow x$. If $x \ge 0$, apply part (ii) of Theorem 10.2, and if $x < 0$, apply part (i); in either case, $F(x_n) \downarrow F(x)$ follows. Thus F is continuous from the right. Finally,

$$(12.5) \qquad \mu(a, b] = F(b) - F(a)$$

for every bounded interval $(a, b]$. If μ is Lebesgue measure, then (12.4) gives $F(x) = x$.

The finite intervals form a π-system generating $\mathscr{R}^1$, and therefore by Theorem 10.3 the function F completely determines μ through the relation (12.5). But (12.5) and μ do not determine F: if $F(x)$ satisfies (12.5), then so does $F(x) + c$. On the other hand, for a given μ, (12.5) certainly determines F to within such an additive constant.

For finite μ, it is customary to standardize F by defining it not by (12.4) but by

$$(12.6) \qquad F(x) = \mu(-\infty, x];$$

then $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = \mu(R^1)$. If μ is a probability measure, F is called a *distribution function* (the adjective *cumulative* is sometimes added).

Measures μ are often specified by means of the function F. The following theorem ensures that to each F there does exist a μ.

Theorem 12.4. *If F is a nondecreasing, right-continuous real function on the line, there exists on $\mathscr{R}^1$ a unique measure μ satisfying (12.5) for all a and b.*

As noted above, uniqueness is a simple consequence of Theorem 10.3. The proof of existence is almost the same as the construction of Lebesgue measure, the case $F(x) = x$. This proof is not carried through at this point, because it is contained in a parallel, more general construction for k-dimensional space in the next theorem. For a very simple argument establishing Theorem 12.4, see the second proof of Theorem 14.1.

Specifying Measures in R^k

The σ-field $\mathscr{R}^k$ of k-dimensional Borel sets is generated by the class of bounded rectangles

$$(12.7) \qquad A = [\, x\colon a_i < x_i \le b_i, \, i = 1, \ldots, k\,]$$

(Example 10.1). If $I_i = (a_i, b_i]$, A has the form of a Cartesian product

$$(12.8) \qquad A = I_1 \times \cdots \times I_k.$$

Consider the sets of the special form

$$(12.9) \qquad S_x = [\, y\colon y_i \le x_i, \, i = 1, \ldots, k\,];$$

S_x consists of the points "southwest" of $x = (x_1, \ldots, x_k)$; in the case $k = 1$ it is the half-infinite interval $(-\infty, x]$. Now S_x is closed, and (12.7) has the form

$$(12.10) \quad A = S_{(b_1 \quad b_k)} - \left[\, S_{(a_1 b_2 \, .. b_k)} \cup S_{(b_1 a_2 \quad b_k)} \cup \cdots \cup S_{(b_1 b_2 \quad a_k)}\,\right].$$

Therefore, the class of sets (12.9) generates $\mathscr{R}^k$. This class is a π-system.

The objective is to find a version of Theorem 12.4 for k-space. This will in particular give k-dimensional Lebesgue measure. The first problem is to find the analogue of (12.5).

A bounded rectangle (12.7) has 2^k vertices—the points $x = (x_1, \ldots, x_k)$ for which each x_i is either a_i or b_i. Let $\mathrm{sgn}_A x$, the signum of the vertex, be $+1$ or -1, according as the number of i $(1 \le i \le k)$ satisfying $x_i = a_i$ is even or odd. For a real function F on R^k, the difference of F around the vertices of A is $\Delta_A F = \Sigma \, \mathrm{sgn}_A x \cdot F(x)$, the sum extending over the 2^k vertices x of A. In the case $k = 1$, $A = (a, b]$ and $\Delta_A F = F(b) - F(a)$. In the case $k = 2$, $\Delta_A F = F(b_1, b_2) - F(b_1, a_2) - F(a_1, b_2) + F(a_1, a_2)$.

Since the k-dimensional analogue of (12.4) is complicated, suppose at first that μ is a finite measure on $\mathscr{R}^k$ and consider instead the analogue of (12.6), namely

(12.11) $F(x) = \mu[\, y: \, y_i \le x_i, \, i = 1, \ldots, k\,].$

Suppose that S_x is defined by (12.9) and A is a bounded rectangle (12.7). Then

(12.12) $\mu(A) = \Delta_A F.$

To see this, apply to the union on the right in (12.10) the inclusion-exclusion formula (10.5). The k sets in the union give $2^k - 1$ intersections, and these are the sets S_x for x ranging over the vertices of A other than $(b_1, \ldots, b_k)$. Taking into account the signs in (10.5) leads to (12.12).
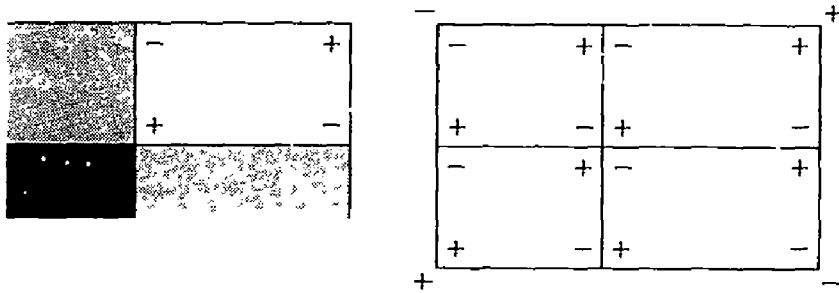
Suppose $x^{(n)} \downarrow x$ in the sense that $x_i^{(n)} \downarrow x_i$ as $n \to \infty$ for each $i = 1, \ldots, k$. Then $S_{x^{(n)}} \downarrow S_x$, and hence $F(x^{(n)}) \to F(x)$ by Theorem 10.2(ii). In this sense, F is *continuous from above*.

Theorem 12.5. *Suppose that the real function F on R^k is continuous from above and satisfies $\Delta_A F \ge 0$ for bounded rectangles A. Then there exists a unique measure μ on $\mathscr{R}^k$ satisfying (12.12) for bounded rectangles A.*

The empty set can be taken as a bounded rectangle (12.7) for which $a_i = b_i$ for some i, and for such a set A, $\Delta_A F = 0$. Thus (12.12) defines a finite-valued set function μ on the class of bounded rectangles. The point of the theorem is that μ extends uniquely to a measure on $\mathscr{R}^k$. The uniqueness is an immediate consequence of Theorem 10.3, since the bounded rectangles form a π-system generating $\mathscr{R}^k$.

If F is bounded, then μ will be a finite measure. But the theorem does not require that F be bounded. The most important unbounded F is $F(x) = x_1 \cdots x_k$. Here $\Delta_A F = (b_1 - a_1) \cdots (b_k - a_k)$ for A given by (12.7). This is the ordinary volume of A as specified by (12.1). The corresponding measure extended to $\mathscr{R}^k$ is k-dimensional Lebesgue measure as described at the beginning of this section.

PROOF OF THEOREM 12.5. As already observed, the uniqueness of the extension is easy to prove. To prove its existence it will first be shown that μ as defined by (12.12) is finitely additive on the class of bounded rectangles. Suppose that each side $I_i = (a_i, b_i]$ of a bounded rectangle (12.7) is partitioned into n_i subintervals $J_{ij} = (t_{i,j-1}, t_{ij}]$, $j = 1, \ldots, n_i$, where $a_i = t_{i0} < t_{i1} < \cdots < t_{in_i} = b_i$. The $n_1 n_2 \cdots n_k$ rectangles

$$(12.13) \qquad B_{j_1 \cdots j_k} = J_{1j_1} \times \cdots \times J_{kj_k}, \qquad 1 \le j_1 \le n_1, \ldots, 1 \le j_k \le n_k,$$

then partition A Call such a partition *regular*. It will first be shown that μ adds for regular partitions:

$$(12.14) \qquad \mu(A) = \sum_{j_1} \cdots \sum_{j_k} \mu(B_{j_1 \cdots j_k}).$$

The right side of (12.14) is $\sum_B \sum_x \mathrm{sgn}_B x \cdot F(x)$, where the outer sum extends over the rectangles B of the form (12.13) and the inner sum extends over the vertices x of B. Now

$$(12.15) \qquad \sum_B \sum_x \mathrm{sgn}_B x \cdot F(x) = \sum_x F(x) \sum_B \mathrm{sgn}_B x,$$

where on the right the outer sum extends over each x that is a vertex of one or more of the B's, and for fixed x the inner sum extends over the B's of which it is a vertex. Suppose that x is a vertex of one or more of the B's but is not a vertex of A. Then there must be an i such that x_i is neither a_i nor b_i. There may be several such i, but fix on one of them and suppose for notational convenience that it is $i = 1$. Then $x_1 = t_{1j}$ with $0 < j < n_1$. The rectangles (12.13) of which x is a vertex therefore come in pairs $B' = B_{jj_2 \cdots j_k}$ and $B'' = B_{j+1,j_2 \cdots j_k}$, and $\mathrm{sgn}_{B'} x = -\mathrm{sgn}_{B''} x$. Thus the inner sum on the right in (12.15) is 0 if x is not a vertex of A.

On the other hand, if x is a vertex of A as well as of at least one B, then for each i either $x_i = a_i = t_{i0}$ or $x_i = b_i = t_{in_i}$. In this case x is a vertex of only one B of the form (12.13)—the one for which $j_i = 1$ or $j_i = n_i$, according as $x_i = a_i$ or $x_i = b_i$— and $\mathrm{sgn}_B x = \mathrm{sgn}_A x$. Thus the right side of (12.15) reduces to $\Delta_A F$, which proves (12.14).

Now suppose that $A = \bigcup_{u=1}^{n} A_u$, where A is the bounded rectangle (12.8), $A_u = I_{1u} \times \cdots \times I_{ku}$ for $u = 1, \ldots, n$, and the A_u are disjoint. For each i ($1 \le i \le k$), the intervals $I_{i1}, \ldots, I_{in}$ have I_i as their union, although they need not be disjoint. But their endpoints split I_i into disjoint subintervals $J_{i1}, \ldots, J_{in_i}$ such that each I_{iu} is the union of certain of the J_{ij}. The rectangles B of the form (12.13) are a regular
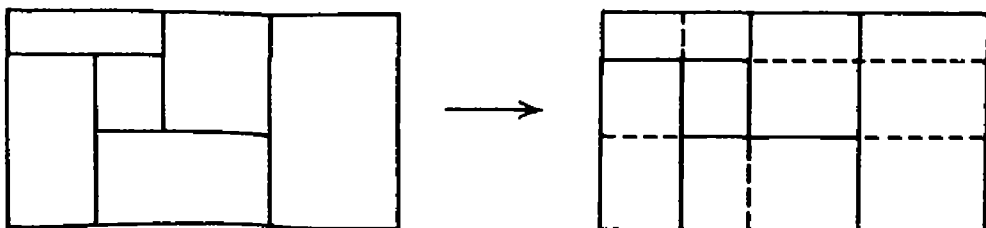
partition of A, as before; furthermore, the B's contained in a single A_u form a regular partition of A_u. Since the A_u are disjoint, it follows by (12.14) that

$$\mu(A) = \sum_B \mu(B) = \sum_{u=1}^{n} \sum_{B \subset A_u} \mu(B) = \sum_{u=1}^{n} \mu(A_u).$$

Therefore, μ is finitely additive on the class $\mathcal{J}^k$ of bounded k-dimensional rectangles.

As shown in Example 11 4, $\mathcal{J}^k$ is a semiring, and so Theorem 11.3 applies. If $A, A_1, \ldots, A_n$ are sets in $\mathcal{J}^k$, then by Lemma 2 of the preceding section,

$$(12.16) \qquad \mu(A) \leq \sum_{u=1}^{n} \mu(A_u) \quad \text{if} \quad A \subset \bigcup_{u=1}^{n} A_u.$$

To apply Theorem 11.3 requires showing that μ is countably subadditive on $\mathcal{J}^k$. Suppose then that $A \subset \bigcup_{u=1}^{\infty} A_u$, where A and the A_u are in $\mathcal{J}^k$. The problem is to prove that

$$(12.17) \qquad \mu(A) \leq \sum_{u=1}^{\infty} \mu(A_u).$$

Suppose that $\epsilon > 0$. If A is given by (12.7) and $B = [x \cdot a_i + \delta < x_i \leq b_i, i \leq k]$, then $\mu(B) > \mu(A) - \epsilon$ for small enough positive δ, because μ is defined by (12.12) and F is continuous from above. Note that A contains the closure $B^- = [x: a_i + \delta \leq x_i \leq b_i, i \leq k]$ of B. Similarly, for each u there is in $\mathcal{J}^k$ a set $B_u = [x: a_{iu} < x_i \leq b_{iu} + \delta_u, i \leq k]$ such that $\mu(B_u) < \mu(A_u) + \epsilon/2^u$ and A_u is in the interior $B_u^{\circ} = [x: a_{iu} < x_i < b_{iu} + \delta_u, i \leq k]$ of B_u.

Since $B^- \subset A \subset \bigcup_{u=1}^{\infty} A_u \subset \bigcup_{u=1}^{\infty} B_u^{\circ}$, it follows by the Heine–Borel theorem that $B \subset B^- \subset \bigcup_{u=1}^{n} B_u^{\circ} \subset \bigcup_{u=1}^{n} B_u$ for some n. Now (12.16) applies, and so $\mu(A) - \epsilon < \mu(B) \leq \sum_{u=1}^{n} \mu(B_u) < \sum_{u=1}^{\infty} \mu(A_u) + \epsilon$. Since ϵ was arbitrary, the proof of (12.17) is complete.

Thus μ as defined by (12.12) is finitely additive and countably subadditive on the semiring $\mathcal{J}^k$. By Theorem 11.3, μ extends to a measure on $\mathcal{R}^k = \sigma(\mathcal{J}^k)$. ∎

Strange Euclidean Sets*

It is possible to construct in the plane a simple curve—the image of $[0, 1]$ under a continuous, one-to-one mapping—having positive area. This is surprising because the curve is *simple*: if the continuous map is not required to be one-to-one, the curve can even fill a square.[†]

Such constructions are counterintuitive, but nothing like one due to Banach and Tarski: Two sets in Euclidean space are congruent if each can be carried onto the other by an isometry, a rigid transformation. Suppose of sets A and B in R^k that A can be decomposed into sets $A_1, \ldots, A_n$ and B can be decomposed into sets $B_1, \ldots, B_n$ in such a way that A_i and B_i are congruent for each $i = 1, \ldots, n$. In this case A and B are said to be *congruent by dissection*. If all the pieces A_i and B_i are Borel sets, then of course $\lambda_k(A) = \sum_{i=1}^{n} \lambda_k(A_i) = \sum_{i=1}^{n} \lambda_k(B_i) = \lambda_k(B)$. But if nonmea-

*This topic may be omitted.
[†]A Peano curve see HAUSDORFF, p 231 For the construction of simple curves of positive area, see GRUBAUM & OLMSTED, pp. 135 ff.

surable sets are allowed in the dissections, then something astonishing happens: *If k ≥ 3, and if A and B are bounded sets in R^k and have nonempty interiors, then A and B are congruent by dissection.* (The result does not hold if k is 1 or 2.)

This is the *Banach-Tarski paradox.* It is usually illustrated in 3-space this way: It is possible to break a solid ball the size of a pea into finitely many pieces and then put them back together again in such a way as to get a solid ball the size of the sun.[†]

PROBLEMS

12.1. Suppose that μ is a measure on $\mathscr{R}^1$ that is finite for bounded sets and is translation-invariant: $\mu(A + x) = u(A)$. Show that $\mu(A) = \alpha\lambda(A)$ for some $\alpha \geq 0$. Extend to R^k

12.2. Suppose that $A \in \mathscr{R}^1$, $\lambda(A) > 0$, and $0 < \theta < 1$. Show that there is a bounded open interval I such that $\lambda(A \cap I) \geq \theta\lambda(I)$. *Hint:* Show that $\lambda(A)$ may be assumed finite, and choose an open G such that $A \subset G$ and $\lambda(A) \geq \theta\lambda(G)$. Now $G = \bigcup_n I_n$ for disjoint open intervals I_n [A12], and $\sum_n \lambda(A \cap I_n) \geq \theta\sum_n \lambda(I_n)$; use an I_n.

12.3. ↑ If $A \in \mathscr{R}^1$ and $\lambda(A) > 0$, then the origin is interior to the difference set $D(A) = [x - y: x, y \in A]$. *Hint* Choose a bounded open interval I as in Problem 12.2 for $\theta = \frac{3}{4}$. Suppose that $|z| < \lambda(I)/2$; since $A \cap I$ and $(A \cap I) + z$ are contained in an interval of length less than $3\lambda(I)/2$ and hence cannot be disjoint, $z \in D(A)$.

12.4. ↑ The following construction leads to a subset H of the unit interval that is nonmeasurable in the extreme sense that its inner and outer Lebesgue measures are 0 and 1: $\lambda_*(H) = 0$ and $\lambda^*(H) = 1$ (see (3.9) and (3.10)). Complete the details. The ideas are those in the construction of a nonmeasurable set at the end of Section 3. It will be convenient to work in $G = [0, 1)$; let $\oplus$ and $\ominus$ denote addition and subtraction modulo 1 in G, which is a group with identity 0.

(a) Fix an irrational θ in G and for $n = 0, \pm 1, \pm 2, \ldots$ let θ_n be $n\theta$ reduced modulo 1. Show that $\theta_n \oplus \theta_m = \theta_{n+m}$, $\theta_n \ominus \theta_m = \theta_{n-m}$, and the θ_n are distinct. Show that $\{\theta_{2n}: n = 0, \pm 1, \ldots\}$ and $\{\theta_{2n+1}: n = 0, \pm 1, \ldots\}$ are dense in G.

(b) Take x and y to be equivalent if $x \ominus y$ lies in $\{\theta_n: n = 0, \pm 1, \ldots\}$, which is a subgroup. Let S contain one representative from each equivalence class (each coset). Show that $G = \bigcup_n(S \oplus \theta_n)$, where the union is disjoint. Put $H = \bigcup_n(S \oplus \theta_{2n})$ and show that $G - H = H \oplus \theta$.

(c) Suppose that A is a Borel set contained in H. If $\lambda(A) > 0$, then $D(A)$ contains an interval $(0, \epsilon)$; but then some θ_{2k+1} lies in $(0, \epsilon) \subset D(A) \subset D(H)$, and so $\theta_{2k+1} = h_1 - h_2 = h_1 \ominus h_2 = (s_1 \oplus \theta_{2n_1}) \ominus (s_2 \oplus \theta_{2n_2})$ for some h_1, h_2 in H and some s_1, s_2 in S. Deduce that $s_1 = s_2$ and obtain a contradiction. Conclude that $\lambda_*(H) = 0$.

(d) Show that $\lambda_*(H \oplus \theta) = 0$ and $\lambda^*(H) = 1$.

12.5. ↑   The construction here gives sets H_n such that $H_n \uparrow G$ and $\lambda_*(H_n) = 0$. If $J_n = G - H_n$, then $J_n \downarrow \varnothing$ and $\lambda^*(J_n) = 1$.

(a) Let $H_n = \bigcup_{k=-n}^{n}(S \oplus \theta_k)$, so that $H_n \uparrow G$. Show that the sets $H_n \oplus \theta_{(2n+1)t}$ are disjoint for different v.

(b) Suppose that A is a Borel set contained in H_n. Show that A and indeed all the $A \oplus \theta_{(2n+1)t}$ have Lebesgue measure 0.

12.6. Suppose that μ is nonnegative and finitely additive on $\mathscr{R}^k$ and that $\mu(R^k) < \infty$. Suppose further that $\mu(A) = \sup \mu(K)$, where K ranges over the compact subsets of A. Show that μ is countably additive (Compare Theorem 12.3(ii).)

12.7. Suppose μ is a measure on $\mathscr{R}^k$ such that bounded sets have finite measure. Given A, show that there exist an F_σ-set U (a countable union of closed sets) and a G_δ-set V (a countable intersection of open sets) such that $U \subset A \subset V$ and $\mu(V - U) = 0$.

12.8. 2.19↑   Suppose that μ is a nonatomic probability measure on $(R^k, \mathscr{R}^k)$ and that $\mu(A) > 0$. Show that there is an uncountable compact set K such that $K \subset A$ and $\mu(K) = 0$.

12.9. The *minimal closed support* of a measure μ on $\mathscr{R}^k$ is a closed set C_μ such that $C_\mu \subset C$ for closed C if and only if C supports μ. Prove its existence and uniqueness. Characterize the points of C_μ as those x such that $\mu(U) > 0$ for every neighborhood U of x. If $k = 1$ and if μ and the function $F(x)$ are related by (12.5), the condition is $F(x - \epsilon) < F(x + \epsilon)$ for all ϵ; x is in this case called a *point of increase* of F.

12.10. Of minor interest is the k-dimensional analogue of (12.4). Let I_t be $(0, t]$ for $t \geq 0$ and $(t, 0]$ for $t \leq 0$, and let $A_x = I_{x_1} \times \cdots \times I_{x_k}$. Let $\varphi(x)$ be $+1$ or -1 according as the number of i, $1 \leq i \leq k$, for which $x_i < 0$ is even or odd. Show that, if $F(x) = \varphi(x)\mu(A_x)$, then (12.12) holds for bounded rectangles A.

Call F degenerate if it is a function of some $k - 1$ of the coordinates, the requirement in the case $k = 1$ being that F is constant. Show that $\Delta_A F = 0$ for every bounded rectangle if and only if F is a finite sum of degenerate functions; (12.12) determines F to within addition of a function of this sort.

12.11. Let G be a nondecreasing, right-continuous function on the line, and put $F(x, y) = \min\{G(x), y\}$. Show that F satisfies the conditions of Theorem 12.5, that the curve $C = \{(x, G(x)): x \in R^1\}$ supports the corresponding measure, and that $\lambda_2(C) = 0$.

12.12. Let F_1 and F_2 be nondecreasing, right-continuous functions on the line and put $F(x_1, x_2) = F_1(x_1)F_2(x_2)$. Show that F satisfies the conditions of Theorem 12.5. Let μ, μ_1, μ_2 be the measures corresponding to F, F_1, F_2, and prove that $\mu(A_1 \times A_2) = \mu_1(A_1)\mu_2(A_2)$ for intervals A_1 and A_2. This μ is the *product* of μ_1 and μ_2; products are studied in a general setting in Section 18.

SECTION 13. MEASURABLE FUNCTIONS AND MAPPINGS

If a real function X on Ω has finite range, it is by the definition in Section 5 a simple random variable if $[\omega\colon X(\omega) = x]$ lies in the basic σ-field $\mathscr{F}$ for each x. The requirement appropriate for the general real function X is stronger; namely, $[\omega\colon X(\omega) \in H]$ must lie in $\mathscr{F}$ for each linear Borel set H. An abstract version of this definition greatly simplifies the theory of such functions.

Measurable Mappings

Let $(\Omega, \mathscr{F})$ and $(\Omega', \mathscr{F}')$ be two measurable spaces. For a mapping $T\colon \Omega \to \Omega'$, consider the inverse images $T^{-1}A' = [\omega \in \Omega\colon T\omega \in A']$ for $A' \subset \Omega'$ (See [A7] for the properties of inverse images.) The mapping T is *measurable* $\mathscr{F}/\mathscr{F}'$ if $T^{-1}A' \in \mathscr{F}$ for each $A' \in \mathscr{F}'$.

For a real function f, the image space Ω' is the line R^1, and in this case $\mathscr{R}^1$ is always tacitly understood to play the role of $\mathscr{F}'$. A real function f on Ω is thus measurable $\mathscr{F}$ (or simply measurable, if it is clear from the context what $\mathscr{F}$ is involved) if it is measurable $\mathscr{F}/\mathscr{R}^1$—that is, if $f^{-1}H = [\omega\colon f(\omega) \in H] \in \mathscr{F}$ for every $H \in \mathscr{R}^1$. In probability contexts, a real measurable function is called a *random variable*. The point of the definition is to ensure that $[\omega\colon f(\omega) \in H]$ has a measure or probability for all sufficiently regular sets H of real numbers—that is, for all Borel sets H.

Example 13.1. A real function f with finite range is measurable if $f^{-1}\{x\} \in \mathscr{F}$ for *each singleton* $\{x\}$, but his is too weak a condition to impose on the general f. (It is satisfied if $(\Omega, \mathscr{F}) = (R^1, \mathscr{R}^1)$ and f is any one-to-one map of the line into itself; but in this case $f^{-1}H$, even for so simple a set H as an interval, can for an appropriately chosen f be any uncountable set, say the non-Borel set constructed in Section 3.) On the other hand, for a measurable f with finite range, $f^{-1}H \in \mathscr{F}$ for *every* $H \subset R^1$; but this is too stiong a condition to impose on the general f. (For $(\Omega, \mathscr{F}) = (R^1, \mathscr{R}^1)$, even $f(x) \equiv x$ fails to satisfy it.) Notice that nothing is required of fA; it need not lie in $\mathscr{R}^1$ for A in $\mathscr{F}$. ∎

If in addition to $(\Omega, \mathscr{F})$, $(\Omega', \mathscr{F}')$, and the map $T\colon \Omega \to \Omega'$, there is a third measurable space $(\Omega'', \mathscr{F}'')$ and a map $T'\colon \Omega' \to \Omega''$, the composition $T'T = T' \circ T$ is the mapping $\Omega \to \Omega''$ that carries ω to $T'(T(\omega))$.

Theorem 13.1. (i) *If $T^{-1}A' \in \mathscr{F}$ for each $A' \in \mathscr{A}'$ and $\mathscr{A}'$ generates $\mathscr{F}'$, then T is measurable $\mathscr{F}/\mathscr{F}'$.*

(ii) *If T is measurable $\mathscr{F}/\mathscr{F}'$ and T' is measurable $\mathscr{F}'/\mathscr{F}''$ then $T'T$ is measurable $\mathscr{F}/\mathscr{F}''$.*

PROOF.   Since $T^{-1}(\Omega' - A') = \Omega - T^{-1}A'$ and $T^{-1}(\bigcup_n A'_n) = \bigcup_n T^{-1}A'_n$, and since $\mathscr{F}$ is a σ-field in Ω, the class $[A': T^{-1}A' \in \mathscr{F}]$ is a σ-field in Ω'. If this σ-field contains $\mathscr{A}'$, it must also contain $\sigma(\mathscr{A}')$, and (i) follows.

As for (ii), it follows by the hypotheses that $A'' \in \mathscr{F}''$ implies that $(T')^{-1}A'' \in \mathscr{F}'$, which in turn implies that $(T'T)^{-1}A'' = [\omega\colon T'T\omega \in A''] = [\omega\colon T\omega \in (T')^{-1}A''] = T^{-1}((T')^{-1}A'') \in \mathscr{F}$.   ∎

By part (i), if f is a real function such that $[\omega\colon F(\omega) \leq x]$ lies in $\mathscr{F}$ for all x, then f is measurable $\mathscr{F}$. This condition is usually easy to check.

Mappings into R^k

For a mapping $f\colon \Omega \to R^k$ carrying Ω into k-space, $\mathscr{R}^k$ is always understood to be the σ-field in the image space. In probabilistic contexts, a measurable mapping into R^k is called a *random vector*. Now f must have the form

$$(13.1) \qquad f(\omega) = (f_1(\omega), \ldots, f_k(\omega))$$

for real functions $f_j(\omega)$. Since the sets (12.9) (the "southwest regions") generate $\mathscr{R}^k$, Theorem 13.1(i) implies that f is measurable $\mathscr{F}$ if and only if the set

$$(13.2) \qquad [\omega\colon f_1(\omega) \leq x_1, \ldots, f_k(\omega) \leq x_k] = \bigcap_{j=1}^{k} [\omega\colon f_j(\omega) \leq x_j]$$

lies in $\mathscr{F}$ for each $(x_1, \ldots, x_k)$. This condition holds if each f_j is measurable $\mathscr{F}$. On the other hand, if $x_j = x$ is fixed and $x_1 = \cdots = x_{j-1} = x_{j+1} = \cdots = x_k = n$ goes to ∞, the sets (13.2) increase to $[\omega\colon f_j(\omega) \leq x]$; the condition thus implies that each f_j is measurable. Therefore, *f is measurable $\mathscr{F}$ if and only if each component function f_j is measurable $\mathscr{F}$*. This provides a practical criterion for mappings into R^k.

A mapping $f\colon R^i \to R^k$ is defined to be measurable if it is measurable $\mathscr{R}^i/\mathscr{R}^k$. Such functions are often called *Borel functions*. To sum up, $T\colon \Omega \to \Omega'$ is measurable $\mathscr{F}/\mathscr{F}'$ if $T^{-1}A' \in \mathscr{F}$ for all $A' \in \mathscr{F}'$; $f\colon \Omega \to R^k$ is measurable $\mathscr{F}$ if it is measurable $\mathscr{F}/\mathscr{R}^k$; and $f\colon R^i \to R^k$ is measurable (a Borel function) if it is measurable $\mathscr{R}^i/\mathscr{R}^k$. If H lies outside $\mathscr{R}^1$, then I_H ($i = k = 1$) is not a Borel function.

Theorem 13.2.   *If $f\colon R^i \to R^k$ is continuous, then it is measurable.*

PROOF.   As noted above, it suffices to check that each set (13.2) lies in $\mathscr{R}^i$. But each is closed because of continuity.   ∎

Theorem 13.3. *If* f_j: $\Omega \to R^1$ *is measurable* $\mathscr{F}$, $j = 1, \ldots, k$, *then* $g(f_1(\omega), \ldots, f_k(\omega))$ *is measurable* $\mathscr{F}$ *if* g: $R^k \to R^1$ *is measurable—in particular, if it is continuous.*

PROOF. If the f_j are measurable, then so is (13.1), so that the result follows by Theorem 13.1(ii). ∎

Taking $g(x_1, \ldots, x_k)$ to be $\sum_{i=1}^{k} x_i$, $\prod_{i=1}^{k} x_i$, and $\max\{x_1, \ldots, x_k\}$ in turn shows that *sums, products, and maxima of measurable functions are measurable.* If $f(\omega)$ is real and measurable, then so are $\sin f(\omega)$, $e^{tf(\omega)}$, and so on, and if $f(\omega)$ never vanishes, then $1/f(\omega)$ is measurable as well.

Limits and Measurability

For a real function f it is often convenient to admit the artificial values ∞ and $-\infty$—to work with the *extended real line* $[-\infty, \infty]$. Such an f is by definition measurable $\mathscr{F}$ if $[\omega: f(\omega) \in H]$ lies in $\mathscr{F}$ for each Borel set H of (finite) real numbers and if $[\omega: f(\omega) = \infty]$ and $[\omega: f(\omega) = -\infty]$ both lie in $\mathscr{F}$. This extension of the notion of measurability is convenient in connection with limits and suprema, which need not be finite.

Theorem 13.4. *Suppose that* $f_1, f_2, \ldots$ *are real functions measurable* $\mathscr{F}$.

(i) *The functions* $\sup_n f_n$, $\inf_n f_n$, $\limsup_n f_n$, *and* $\liminf_n f_n$ *are measurable* $\mathscr{F}$.

(ii) *If* $\lim_n f_n$ *exists everywhere, then it is measurable* $\mathscr{F}$.

(iii) *The* ω-*set where* $\{f_n(\omega)\}$ *converges lies in* $\mathscr{F}$.

(iv) *If* f *is measurable* $\mathscr{F}$, *then the* ω-*set where* $f_n(\omega) \to f(\omega)$ *lies in* $\mathscr{F}$.

PROOF. Clearly, $[\sup_n f_n \leq x] = \bigcap_n [f_n \leq x]$ lies in $\mathscr{F}$ even for $x = \infty$ and $x = -\infty$, and so $\sup_n f_n$ is measurable. The measurability of $\inf_n f_n$ follows the same way, and hence $\limsup_n f_n = \inf_n \sup_{k \geq n} f_k$ and $\liminf_n f_n = \sup_n \inf_{k \geq n} f_k$ are measurable. If $\lim_n f_n$ exists, it coincides with these last two functions and hence is measurable. Finally, the set in (iii) is the set where $\limsup_n f_n(\omega) = \liminf_n f_n(\omega)$, and that in (iv) is the set where this common value is $f(\omega)$. ∎

Special cases of this theorem have been encountered before—part (iv), for example, in connection with the strong law of large numbers. The last three parts of the theorem obviously carry over to mappings into R^k.

A *simple* real function is one with finite range; it can be put in the form

$$(13.3) \qquad\qquad f = \sum x_i I_{A_i},$$

where the A_i form a finite decomposition of Ω. It is measurable $\mathscr{F}$ if each A_i lies in $\mathscr{F}$. The simple random variables of Section 5 have this form.

Many results concerning measurable functions are most easily proved first for simple functions and then, by an appeal to the next theorem and a passage to the limit, for the general measurable function.

Theorem 13.5. *If f is real and measurable $\mathscr{F}$, there exists a sequence $\{f_n\}$ of simple functions, each measurable $\mathscr{F}$, such that*

$$(13.4) \qquad\qquad 0 \le f_n(\omega) \uparrow f(\omega) \qquad iff(\omega) \ge 0$$

and

$$(13.5) \qquad\qquad 0 \ge f_n(\omega) \downarrow f(\omega) \qquad iff(\omega) \le 0.$$

Proof. Define

$$(13.6) \quad f_n(\omega) = \begin{cases} -n & \text{if } -\infty \le f(\omega) \le -n, \\ -(k-1)2^{-n} & \text{if } -k2^{-n} < f(\omega) \le -(k-1)2^{-n}, \\ & \qquad 1 \le k \le n2^n, \\ (k-1)2^{-n} & \text{if } (k-1)2^{-n} \le f(\omega) < k2^{-n}, \\ & \qquad 1 \le k \le n2^n, \\ n & \text{if } n \le f(\omega) \le \infty. \end{cases}$$

This sequence has the required properties.   ∎

Note that (13.6) covers the possibilities $f(\omega) = \infty$ and $f(\omega) = -\infty$.

If $A \in \mathscr{F}$, a function f defined only on A is by definition measurable if $[\omega \in A: f(\omega) \in H]$ lies in $\mathscr{F}$ for $H \in \mathscr{R}^1$ and for $H = \{\infty\}$ and $H = \{-\infty\}$.

Transformations of Measures

Let $(\Omega, \mathscr{F})$ and $(\Omega', \mathscr{F}')$ be measurable spaces, and suppose that the mapping $T: \Omega \to \Omega'$ is measurable $\mathscr{F}/\mathscr{F}'$. Given a measure μ on $\mathscr{F}$, define a set function μT^{-1} on $\mathscr{F}'$ by

$$(13.7) \qquad\qquad \mu T^{-1}(A') = \mu(T^{-1}A'), \qquad A' \in \mathscr{F}'.$$

That is, μT^{-1} assigns value $\mu(T^{-1}A')$ to the set A'. If $A' \in \mathscr{F}'$, then $T^{-1}A' \in \mathscr{F}$ because T is measurable, and hence the set function μT^{-1} is well defined on $\mathscr{F}'$. Since $T^{-1} \bigcup_n A'_n = \bigcup_n T^{-1}A'_n$ and the $T^{-1}A'_n$ are disjoint

sets in Ω if the A'_n are disjoint sets in Ω', the countable additivity of μT^{-1} follows from that of μ. Thus μT^{-1} *is a measure.* This way of transferring a measure from Ω to Ω' will prove useful in a number of ways.

If μ is finite, so is μT^{-1}; if μ is a probability measure, so is μT^{-1}.[†]

PROBLEMS

13.1. Functions are often defined in pieces (for example, let $f(x)$ be x^3 or x^{-1} as $x \geq 0$ or $x < 0$), and the following result shows that the function is measurable if the pieces are.

Consider measurable spaces $(\Omega, \mathscr{F})$ and $(\Omega', \mathscr{F}')$ and a map $T \; \Omega \to \Omega'$ Let $A_1, A_2, \ldots$ be a countable covering of Ω by $\mathscr{F}$-sets. Consider the σ-field $\mathscr{F}_n = [A: A \subset A_n, A \in \mathscr{F}]$ in A_n and the restriction T_n of T to A_n. Show that T is measurable $\mathscr{F}/\mathscr{F}'$ if and only if T_n is measurable $\mathscr{F}_n/\mathscr{F}'$ for each n.

13.2. (a) For a map T and σ-fields $\mathscr{F}$ and $\mathscr{F}'$, define $T^{-1}\mathscr{F}' = [T^{-1}A': A' \in \mathscr{F}']$ and $T\mathscr{F} = [A': T^{-1}A' \in \mathscr{F}]$. Show that $T^{-1}\mathscr{F}'$ and $T\mathscr{F}$ are σ-fields and that measurability $\mathscr{F}/\mathscr{F}'$ is equivalent to $T^{-1}\mathscr{F}' \subset \mathscr{F}$ and to $\mathscr{F}' \subset T\mathscr{F}$.

(b) For given $\mathscr{F}'$, $T^{-1}\mathscr{F}'$, which is the smallest σ-field for which T is measurable $\mathscr{F}/\mathscr{F}'$, is by definition the σ-field *generated* by T. For simple random variables describe $\sigma(X_1, \ldots, X_n)$ in these terms.

(c) Let $\sigma'(\mathscr{A}')$ be the σ-field in Ω' generated by $\mathscr{A}'$. Show that $\sigma(T^{-1}\mathscr{A}') = T^{-1}(\sigma'(\mathscr{A}'))$. Prove Theorem 10.1 by taking T to be the identity map from Ω_0 to Ω.

13.3. ↑ Suppose that $f: \Omega \to R^1$. Show that f is measurable $T^{-1}\mathscr{F}'$ if and only if there exists a map $\varphi: \Omega' \to R^1$ such that φ is measurable $\mathscr{F}'$ and $f = \varphi T$. *Hint:* First consider simple functions and then use Theorem 13.5.

13.4. ↑ Relate the result in Problem 13.3 to Theorem 5.1(ii).

13.5. Show of real functions f and g that $f(\omega) + g(\omega) < x$ if and only if there exist rationals r and s such that $r + s < x$, $f(\omega) < r$, and $g(\omega) < s$. Prove directly that $f + g$ is measurable $\mathscr{F}$ if f and g are.

13.6. Let $\mathscr{F}$ be a σ-field in R^1. Show that $\mathscr{R}^1 \subset \mathscr{F}$ if and only if every continuous function is measurable $\mathscr{F}$. Thus $\mathscr{R}^1$ is the smallest σ-field with respect to which all the continuous functions are measurable.

13.7. Consider on R^1 the smallest class $\mathscr{X}$ (that is, the intersection of all classes) of real functions containing all the continuous functions and closed under pointwise passages to the limit. The elements of $\mathscr{X}$ are called *Baire* functions. Show that Baire functions and Borel functions on R^1 are the same thing.

13.8. A real function f on the line is upper semicontinuous at x if for each ϵ there is a δ such that $|x - y| < \delta$ implies that $f(y) < f(x) + \epsilon$. Show that, if f is everywhere upper semicontinuous, then it is measurable.

[†]But see Problem 13.14.

13.9. Suppose that f_n and f are finite-valued, $\mathscr{F}$-measurable functions such that $f_n(\omega) \to f(\omega)$ for $\omega \in A$, where $\mu(A) < \infty$ (μ a measure on $\mathscr{F}$). Prove *Egoroff's theorem*: For each ϵ there exists a subset B of A such that $\mu(B) < \epsilon$ and $f_n(\omega) \to f(\omega)$ uniformly on $A - B$. *Hint:* Let $B_i^{(k)}$ be the set of ω in A such that $|f(\omega) - f_i(\omega)| > k^{-1}$ for some $i \geq n$. Show that $B_n^{(k)} \downarrow / \varnothing$ as $n \uparrow \infty$, choose n_k so that $\mu(B_{n_k}^{(k)}) < \epsilon/2^k$, and put $B = \bigcup_{k=1}^{\infty} B_{n_k}^{(k)}$.

13.10. ↑   Show that Egoroff's theorem is false without the hypothesis $\mu(A) < \infty$.

13.11. 2.9↑   Show that, if f is measurable $\sigma(\mathscr{A})$, then there exists a countable subclass $\mathscr{A}_f$ of $\mathscr{A}$ such that f is measurable $\sigma(\mathscr{A}_f)$.

13.12. *Circular Lebesgue measure*   Let C be the unit circle in the complex plane, and define $T: [0, 1) \to C$ by $T\omega = e^{2\pi i \omega}$. Let $\mathscr{B}$ consist of the Borel subsets of $[0, 1)$, and let λ be Lebesgue measure on $\mathscr{B}$. Show that $\mathscr{C} = [A: T^{-1}A \in \mathscr{B}]$ consists of the sets in $\mathscr{R}^2$ (identify R^2 with the complex plane) that are contained in C Show that $\mathscr{C}$ is generated by the arcs of C. Circular Lebesgue measure is defined as $\mu = \lambda T^{-1}$. Show that μ is invariant under rotations: $\mu[\theta z: z \in A] = \mu(A)$ for $A \in \mathscr{C}$ and $\theta \in C$.

13.13. ↑   Suppose that the circular Lebesgue measure of A satisfies $\mu(A) > 1 - n^{-1}$ and that B contains at most n points. Show that some rotation carries B into A $\theta B \subset A$ foı some θ in C.

13.14. Show by example that μ σ-finite does not imply μT^{-1} σ-finite.

13.15. Consider Lebesgue measure λ restricted to the class $\mathscr{B}$ of Borel sets in $(0, 1]$. For a fixed permutation $n_1, n_2, \ldots$ of the positive integers, if x has dyadic expansion $.x_1 x_2 \ldots$, take $Tx = .x_{n_1} x_{n_2} \ldots$. Show that T is measurable $\mathscr{B}/\mathscr{B}$ and that $\lambda T^{-1} = \lambda$.

13.16. Let H_k be the union of the intervals $((i - 1)/2^k, i/2^k]$ for i even, $1 \leq i \leq 2^k$. Show that if $0 < f(\omega) \leq 1$ for all ω and $A_k = f^{-1}(H_k)$, then $f(\omega) = \sum_{k=1}^{\infty} I_{A_k}(\omega)/2^k$, an infinite linear combination of indicators.

13.17. Let $S = \{0, 1\}$, and define a map T from sequence space S^{∞} to $[0, 1]$ by $T\omega = \sum_{k=1}^{\infty} a_k(\omega)/2^k$. Define a map U of $[0, 1]$ to S^{∞} by $Ux = (d_1(x), d_2(x), \ldots)$, where the $d_k(x)$ are the digits of the nonterminating dyadic expansion of x (and $d_k(0) \equiv 0$). Show that T is measurable $\mathscr{C}/\mathscr{B}$ and that U is measurable $\mathscr{B}/\mathscr{C}$. Let P be the measure specified by (2.21) for $p_0 = p_1 = \frac{1}{2}$. Describe PT^{-1} and λU^{-1}.

SECTION 14. DISTRIBUTION FUNCTIONS

Distribution Functions

A random variable as defined in Section 13 is a measurable real function X on a probability measure space $(\Omega, \mathscr{F}, P)$. The *distribution* or *law* of the

random variable is the probability measure μ on $(R^1, \mathscr{R}^1)$ defined by

$$(14.1) \qquad\qquad \mu(A) = P[X \in A], \qquad A \in \mathscr{R}^1.$$

As in the case of the simple random variables in Chapter 1, the argument ω is usually omitted: $P[X \in A]$ is short for $P[\omega : X(\omega) \in A]$. In the notation (13.7), the distribution is PX^{-1}.

For simple random variables the distribution was defined in Section 5—see (5.12). There μ was defined for every subset of the line, however; from now on μ will be defined only for Borel sets, because unless X is simple, one cannot in general be sure that $[X \in A]$ has a probability for A outside $\mathscr{R}^1$.

The *distribution function* of X is defined by

$$(14.2) \qquad\qquad F(x) = \mu(-\infty, x] = P[X \le x]$$

for real x. By continuity from above (Theorem 10.2(ii)) for μ, F is right-continuous. Since F is nondecreasing, the left-hand limit $F(x-) = \lim_{y \uparrow x} F(y)$ exists, and by continuity from below (Theorem 10.2(i)) for μ,

$$(14.3) \qquad\qquad F(x-) = \mu(-\infty, x) = P[X < x].$$

Thus the jump or saltus in F at x is

$$F(x) - F(x-) = \mu\{x\} = P[X = x].$$

Therefore (Theorem 10.2(iv)) F can have at most countably many points of discontinuity. Clearly,

$$(14.4) \qquad\qquad \lim_{x \to -\infty} F(x) = 0, \qquad \lim_{x \to \infty} F(x) = 1.$$

A function with these properties must, in fact, be the distribution function of some random variable:

Theorem 14.1. *If F is a nondecreasing, right-continuous function satisfying* (14.4), *then there exists on some probability space a random variable X for which $F(x) = P[X \le x]$.*

FIRST PROOF.   By Theorem 12.4, if F is nondecreasing and right-continuous, there is on $(R^1, \mathscr{R}^1)$ a measure μ for which $\mu(a, b] = F(b) - F(a)$. But $\lim_{x \to -\infty} F(x) = 0$ implies that $\mu(-\infty, x] = F(x)$, and $\lim_{x \to \infty} F(x) = 1$ implies that $\mu(R^1) = 1$. For the probability space take $(\Omega, \mathscr{F}, P) = (R^1, \mathscr{R}^1, \mu)$, and for X take the identity function: $X(\omega) \equiv \omega$. Then $P[X \le x] = \mu[\omega \in R^1 : \omega \le x] = F(x)$. ∎

SECOND PROOF. There is a proof that uses only the existence of Lebesgue measure on the unit interval and does not require Theorem 12.4. For the probability space take the open unit interval: Ω is $(0, 1)$, $\mathscr{F}$ consists of the Borel subsets of $(0, 1)$, and $P(A)$ is the Lebesgue measure of A.

To understand the method, suppose at first that F is continuous and strictly increasing. Then F is a one-to-one mapping of R^1 onto $(0,1)$; let φ: $(0, 1) \to R^1$ be the inverse mapping. For $0 < \omega < 1$, let $X(\omega) = \varphi(\omega)$. Since φ is increasing, certainly X is measurable $\mathscr{F}$. If $0 < u < 1$, then $\varphi(u) \leq x$ if and only if $u \leq F(x)$. Since P is Lebesgue measure, $P[X \leq x] = P[\omega \in (0,1):$ $\varphi(\omega) \leq x] = P[\omega \in (0, 1): \omega \leq F(x)] = F(x)$, as required.
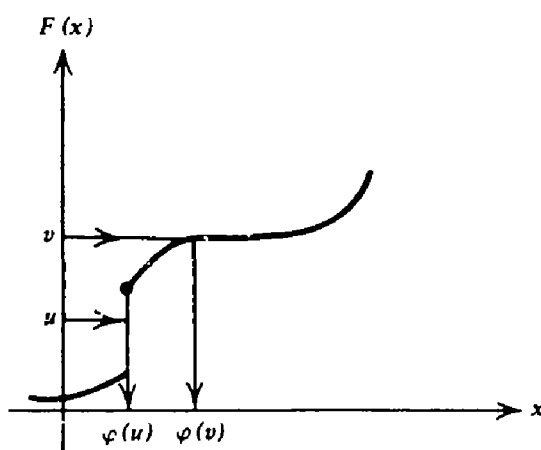
If F has discontinuities or is not strictly increasing, define[†]

$$(14.5) \qquad\qquad \varphi(u) = \inf[x : u \leq F(x)]$$

for $0 < u < 1$. Since F is nondecreasing, $[x: u \leq F(x)]$ is an interval stretching to ∞; since F is right-continuous, this interval is closed on the left. For $0 < u < 1$, therefore, $[x: u \leq F(x)] = [\varphi(u), \infty)$, and so $\varphi(u) \leq x$ if and only if $u \leq F(x)$. If $X(\omega) = \varphi(\omega)$ for $0 < \omega < 1$, then by the same reasoning as before, X is a random variable and $P[X \leq x] = F(x)$. ∎

This second argument actually provides a simple proof of Theorem 12.4 for a probability distribution[‡] F: the distribution μ (as defined by (14.1)) of the random variable just constructed satisfies $\mu(-\infty, x] = F(x)$ and hence $\mu(a, b] = F(b) - F(a)$.

Exponential Distributions

There are a number of results which for their interpretation require random variables, independence, and other probabilistic concepts, but which can be discussed technically in terms of distribution functions alone and do not require the apparatus of measure theory.

[†]This is called the quantile function
[‡]For the general case, see Problem 14.2.

Suppose as an example that F is the distribution function of the waiting time to the occurrence of some event—say the arrival of the next customer at a queue or the next call at a telephone exchange. As the waiting time must be positive, assume that $F(0) = 0$. Suppose that $F(x) < 1$ for all x, and furthermore suppose that

$$(14.6) \qquad \frac{1 - F(x+y)}{1 - F(x)} = 1 - F(y), \qquad x, y \geq 0.$$

The right side of this equation is the probability that the waiting time exceeds y; by the definition (4.1) of conditional probability, the left side is the probability that the waiting time exceeds $x + y$ given that it exceeds x. Thus (14.6) attributes to the waiting-time mechanism a kind of lack of memory or aftereffect: If after a lapse of x units of time the event has not yet occurred, the waiting time still remaining is conditionally distributed just as the entire waiting time from the beginning. For reasons that will emerge later (see Section 23), waiting times often have this property.

The condition (14.6) completely determines the form of F. If $U(x) = 1 - F(x)$, (14.6) is $U(x+y) = U(x)U(y)$. This is a form of Cauchy's equation [A20], and since U is bounded, $U(x) = e^{-\alpha x}$ for some α. Since $\lim_{x \to \infty} U(x) = 0$, α must be positive. Thus (14.6) implies that F has the exponential form

$$(14.7) \qquad F(x) = \begin{cases} 0 & \text{if } x \leq 0, \\ 1 - e^{-\alpha x} & \text{if } x \geq 0, \end{cases}$$

and conversely.

Weak Convergence

Random variables $X_1, \ldots, X_n$ are defined to be independent if the events $[X_1 \in A_1], \ldots, [X_n \in A_n]$ are independent for all Borel sets $A_1, \ldots, A_n$, so that $P[X_i \in A_i, i = 1, \ldots, n] = \prod_{i=1}^{n} P[X_i \in A_i]$. To find the distribution function of the maximum $M_n = \max\{X_1, \ldots, X_n\}$, take $A_1 = \cdots = A_n = (-\infty, x]$. This gives $P[M_n \leq x] = \prod_{i=1}^{n} P[X_i \leq x]$. If the X_i are independent and have common distribution function G and M_n has distribution function F_n, then

$$(14.8) \qquad F_n(x) = G^n(x).$$

It is possible without any appeal to measure theory to study the real function F_n solely by means of the relation (14.8), which can indeed be taken as defining F_n. It is possible in particular to study the asymptotic properties of F_n:

Example 14.1. Consider a stream or sequence of events, say arrivals of calls at a telephone exchange. Suppose that the times between successive events, the interarrival times, are independent and that each has the exponential form (14.7) with a common value of α. By (14.8) the maximum M_n among the first n interarrival times has distribution function $F_n(x) = (1 - e^{-\alpha x})^n$, $x \geq 0$. For each x, $\lim_n F_n(x) = 0$, which means that M_n tends to be large for n large. But $P[M_n - \alpha^{-1} \log n \leq x] = F_n(x + \alpha^{-1} \log n)$. This is the distribution function of $M_n - \alpha^{-1} \log n$, and it satisfies

$$(14.9) \qquad F_n(x + \alpha^{-1} \log n) = (1 - e^{-(\alpha x + \log n)})^n \to e^{-e^{-\alpha x}}$$

as $n \to \infty$; the equality here holds if $\log n \geq -\alpha x$, and so the limit holds for all x. This gives for large n the approximate distribution of the normalized random variable $M_n - \alpha^{-1} \log n$. ∎

If F_n and F are distribution functions, then by definition, F_n *converges weakly* to F, written $F_n \Rightarrow F$, if

$$(14.10) \qquad \lim_n F_n(x) = F(x)$$

for each x at which F is continuous.[†] To study the approximate distribution of a random variable Y_n it is often necessary to study instead the normalized or rescaled random variable $(Y_n - b_n)/a_n$ for appropriate constants a_n and b_n. If Y_n has distribution function F_n and if $a_n > 0$, then $P[(Y_n - b_n)/a_n \leq x] = P[Y_n \leq a_n x + b_n]$, and therefore $(Y_n - b_n)/a_n$ has distribution function $F_n(a_n x + b_n)$. For this reason weak convergence often appears in the form[‡]

$$(14.11) \qquad F_n(a_n x + b_n) \Rightarrow F(x).$$

An example of this is (14.9): there $a_n = 1$, $b_n = \alpha^{-1} \log n$, and $F(x) = e^{-e^{-\alpha x}}$.

Example 14.2. Consider again the distribution function (14.8) of the maximum, but suppose that G has the form

$$G(x) = \begin{cases} 0 & \text{if } x \leq 1, \\ 1 - x^{-\alpha} & \text{if } x \geq 1, \end{cases}$$

where $\alpha > 0$. Here $F_n(n^{1/\alpha} x) = (1 - n^{-1} x^{-\alpha})^n$ for $x \geq n^{-1/\alpha}$, and therefore

$$\lim_n F_n(n^{1/\alpha} x) = \begin{cases} 0 & \text{if } x \leq 0, \\ e^{-x^{-\alpha}} & \text{if } x > 0. \end{cases}$$

This is an example of (14.11) in which $a_n = n^{1/\alpha}$ and $b_n = 0$. ∎

[†]For the role of continuity, see Example 14.4.

[‡]To write $F_n(a_n x + b_n) \Rightarrow F(x)$ ignores the distinction between a function and its value at an unspecified value of its argument, but the meaning of course is that $F_n(a_n x + b_n) \to F(x)$ at continuity points x of F.

Example 14.3. Consider (14.8) once more, but for

$$G(x) = \begin{cases} 0 & \text{if } x \le 0, \\ 1 - (1-x)^\alpha & \text{if } 0 \le x \le 1, \\ 1 & \text{if } x \ge 1, \end{cases}$$

where $\alpha > 0$. This time $F_n(n^{-1/\alpha}x + 1) = (1 - n^{-1}(-x)^\alpha)^n$ if $-n^{1/\alpha} \le x \le 0$. Therefore,

$$\lim_n F_n(n^{-1/\alpha}x + 1) = \begin{cases} e^{-(-x)^\alpha} & \text{if } x \le 0, \\ 1 & \text{if } x > 0, \end{cases}$$

a case of (14.11) in which $a_n = n^{-1/\alpha}$ and $b_n = 1$.   ∎

Let Δ be the distribution function with a unit jump at the origin:

(14.12) $\Delta(x) = \begin{cases} 0 & \text{if } x < 0, \\ 1 & \text{if } x \ge 0. \end{cases}$

If $X(\omega) \equiv 0$, then X has distribution function Δ.

Example 14.4. Let $X_1 X_2, \ldots$ be independent random variables for which $P[X_k = 1] = P[X_k = -1] = \frac{1}{2}$, and put $S_n = X_1 + \cdots + X_n$. By the weak law of large numbers,

(14.13) $P\left[\left|n^{-1}S_n\right| > \epsilon\right] \to 0$

for $\epsilon > 0$. Let F_n be the distribution function of $n^{-1}S_n$. If $x > 0$, then $F_n(x) = 1 - P[n^{-1}S_n > x] \to 1$; if $x < 0$, then $F_n(x) \le P[|n^{-1}S_n| \ge |x|] \to 0$. As this accounts for all the continuity points of Δ, $F_n \Rightarrow \Delta$. It is easy to turn the argument around and deduce (14.13) from $F_n \Rightarrow \Delta$. Thus the weak law of large numbers is equivalent to the assertion that the distribution function of $n^{-1}S_n$ converges weakly to Δ.

If n is odd, so that $S_n = 0$ is impossible, then by symmetry the events $[S_n \le 0]$ and $[S_n \ge 0]$ each have probability $\frac{1}{2}$ and hence $F_n(0) = \frac{1}{2}$. Thus $F_n(0)$ does not converge to $\Delta(0) = 1$, but because Δ is discontinuous at 0, the definition of weak convergence does not require this.   ∎

Allowing (14.10) to fail at discontinuity points x of F thus makes it possible to bring the weak law of large numbers under the theory of weak convergence. But if (14.10) need hold only for certain values of x, there

arises the question of whether weak limits are unique. Suppose that $F_n \Rightarrow F$ and $F_n \Rightarrow G$. Then $F(x) = \lim_n F_n(x) = G(x)$ if F and G are both continuous at x. Since F and G each have only countably many points of discontinuity,[†] the set of common continuity points is dense, and it follows by right continuity that F and G are identical. A sequence can thus have at most one weak limit.

Convergence of distribution functions is studied in detail in Chapter 5. The remainder of this section is devoted to some weak-convergence theorems which are interesting both for themselves and for the reason that they require so little technical machinery.

Convergence of Types[*]

Distribution functions F and G are of the same *type* if there exist constants a and b, $a > 0$, such that $F(ax + b) = G(x)$ for all x. A distribution function is *degenerate* if it has the form $\Delta(x - x_0)$ (see (14.12)) for some x_0; otherwise, it is *nondegenerate*.

Theorem 14.2. *Suppose that* $F_n(u_n x + v_n) \Rightarrow F(x)$ *and* $F_n(a_n x + b_n) \Rightarrow G(x)$, *where* $u_n > 0$, $a_n > 0$, *and* F *and* G *are nondegenerate. Then there exist* a *and* b, $a > 0$, *such that* $a_n/u_n \to a$, $(b_n - v_n)/u_n \to b$, *and* $F(ax + b) = G(x)$.

Thus there can be only one possible limit type and essentially only one possible sequence of norming constants.

The proof of the theorem is for clarity set out in a sequence of lemmas. In all of them, a and the a_n are assumed to be positive.

Lemma 1. *If* $F_n \Rightarrow F$, $a_n \to a$, *and* $b_n \to b$, *then* $F_n(a_n x + b_n) \Rightarrow F(ax + b)$.

PROOF. If x is a continuity point of $F(ax + b)$ and $\epsilon > 0$, choose continuity points u and v of F so that $u < ax + b < v$ and $F(v) - F(u) < \epsilon$; this is possible because F has only countably many discontinuities. For large enough n, $u < a_n x + b_n < v$, $|F_n(u) - F(u)| < \epsilon$, and $|F_n(v) - F(v)| < \epsilon$; but then $F(ax + b) - 2\epsilon < F(u) - \epsilon < F_n(u) \le F_n(a_n x + b_n) \le F_n(v) < F(v) + \epsilon < F(ax + b) + 2\epsilon$. ∎

Lemma 2. *If* $F_n \Rightarrow F$ *and* $a_n \to \infty$, *then* $F_n(a_n x) \Rightarrow \Delta(x)$.

PROOF. Given ϵ, choose a continuity point u of F so large that $F(u) > 1 - \epsilon$. If $x > 0$, then for all large enough n, $a_n x > u$ and $|F_n(u) - F(u)| < \epsilon$, so

---

[†] The proof following (14.3) uses measure theory, but this is not necessary If the saltus $\sigma(x) = F(x) - F(x-)$ exceeds ϵ àt $x_1 < \cdot \cdot < x_n$, then $F(x_i) - F(x_{i-1}) > \epsilon$ (take $x_0 < x_1$), and so $n\epsilon \le F(x_n) - F(x_0) \le 1$; hence $[x \cdot \sigma(x) > \epsilon]$ is finite and $[x \colon \sigma(x) > 0]$ is countable

[*] This topic may be omitted.

that $F_n(a_n x) \geq F_n(u) > F(u) - \epsilon > 1 - 2\epsilon$. Thus $\lim_n F_n(a_n x) = 1$ for $x > 0$; similarly, $\lim_n F_n(a_n x) = 0$ for $x < 0$.   ∎

Lemma 3. *If $F_n \Rightarrow F$ and b_n is unbounded, then $F_n(x + b_n)$ cannot converge weakly.*

PROOF. Suppose that b_n is unbounded and that $b_n \to \infty$ along some subsequence (the case $b_n \to -\infty$ is similar). Suppose that $F_n(x + b_n) \Rightarrow G(x)$. Given ϵ, choose a continuity point u of F so that $F(u) > 1 - \epsilon$. Whatever x may be, for n far enough out in the subsequence, $x + b_n > u$ and $F_n(u) > 1 - 2\epsilon$, so that $F_n(x + b_n) > 1 - 2\epsilon$. Thus $G(x) = \lim_n F_n(x + b_n) = 1$ for all continuity points x of G, which is impossible.   ∎

Lemma 4. *If $F_n(x) \Rightarrow F(x)$ and $F_n(a_n x + b_n) \Rightarrow G(x)$, where F and G are nondegenerate, then*

$$(14.14) \qquad 0 < \inf_n a_n \leq \sup_n a_n < \infty, \qquad \sup_n |b_n| < \infty.$$

PROOF. Suppose that a_n is not bounded above. Arrange by passing to a subsequence that $a_n \to \infty$. Then by Lemma 2,

$$(14.15) \qquad F_n(a_n x) \Rightarrow \Delta(x).$$

Since

$$(14.16) \qquad F_n\big(a_n(x + b_n/a_n)\big) = F_n(a_n x + b_n) \Rightarrow G(x),$$

it follows by Lemma 3 that b_n/a_n is bounded along this subsequence. By passing to a further subsequence, arrange that b_n/a_n converges to some c. By (14.15) and Lemma 1, $F_n(a_n(x + b_n/a_n)) \Rightarrow \Delta(x + c)$ along this subsequence. But (14.16) now implies that G is degenerate, contrary to hypothesis.

Thus a_n is bounded above. If $G_n(x) = F_n(a_n x + b_n)$, then $G_n(x) \Rightarrow G(x)$ and $G_n(a_n^{-1} x - a_n^{-1} b_n) = F_n(x) \Rightarrow F(x)$. The result just proved shows that a_n^{-1} is bounded.

Thus a_n is bounded away from 0 and ∞. If b_n is not bounded, neither is b_n/a_n; pass to a subsequence along which $b_n/a_n \to \pm\infty$ and a_n converges to a positive a. Since, by Lemma 1, $F_n(a_n x) \Rightarrow F(ax)$ along the subsequence, (14.16) and $b_n/a_n \to \pm\infty$ stand in contradiction (Lemma 3 again). Therefore b_n is bounded.   ∎

Lemma 5. *If $F(x) = F(ax + b)$ for all x and F is nondegenerate, then $a = 1$ and $b = 0$.*

PROOF. Since $F(x) = F(a^n x + (a^{n-1} + \cdots + a + 1)b)$, it follows by Lemma 4 that a^n is bounded away from 0 and ∞, so that $a = 1$, and it then follows that nb is bounded, so that $b = 0$.   ∎

PROOF OF THEOREM 14.2. Suppose first that $u_n = 1$ and $v_n = 0$. Then (14.14) holds. Fix any subsequence along which a_n converges to some positive a and b_n converges to some b. By Lemma 1, $F_n(a_n x + b_n) \Rightarrow F(ax + b)$ along this subsequence, and the hypothesis gives $F(ax + b) = G(x)$.

Suppose that along some other sequence, $a_n \to u > 0$ and $b_n \to v$. Then $F(ux + v) = G(x)$ and $F(ax + b) = G(x)$ both hold, so that $u = a$ and $v = b$ by Lemma 5. Every convergent subsequence of $\{(a_n, b_n)\}$ thus converges to (a, b), and so the entire sequence does.

For the general case, let $H_n(x) = F_n(u_n x + v_n)$. Then $H_n(x) \Rightarrow F(x)$ and $H_n(a_n u_n^{-1} x + (b_n - v_n)u_n^{-1}) \Rightarrow G(x)$, and so by the case already treated, $a_n u_n^{-1}$ converges to some positive a and $(b_n - v_n)u_n^{-1}$ to some b, and as before, $F(ax + b) = G(x)$. ∎

Extremal Distributions*

A distribution function F is *extremal* if it is nondegenerate and if, for some distribution function G and constants a_n $(a_n > 0)$ and b_n,

$$(14.17) \qquad\qquad G^n(a_n x + b_n) \Rightarrow F(x).$$

These are the possible limiting distributions of normalized maxima (see (14.8)), and Examples 14.1, 14.2, and 14.3 give three specimens. The following analysis shows that these three examples exhaust the possible types.

Assume that F is extremal. From (14.17) follow $G^{nk}(a_n x + b_n) \Rightarrow F^k(x)$ and $G^{nk}(a_{nk} x + b_{nk}) \Rightarrow F(x)$, and so by Theorem 14.2 there exist constants c_k and d_k such that c_k is positive and

$$(14.18) \qquad\qquad F^k(x) = F(c_k x + d_k).$$

From $F(c_{jk} x + d_{jk}) = F^{jk}(x) = F^j(c_k x + d_k) = F(c_j(c_k x + d_k) + d_j)$ follow (Lemma 5) the relations

$$(14.19) \qquad\qquad c_{jk} = c_j c_k, \qquad d_{jk} = c_j d_k + d_j = c_k d_j + d_k.$$

Of course, $c_1 = 1$ and $d_1 = 0$. There are three cases to be considered separately.

CASE 1. Suppose that $c_k = 1$ for all k. Then

$$(14.20) \qquad F^k(x) = F(x + d_k), \qquad F^{1/k}(x) = F(x - d_k).$$

This implies that $F^{j/k}(x) = F(x + d_j - d_k)$. For positive rational $r = j/k$, put $\delta_r = d_j - d_k$; (14.19) implies that the definition is consistent, and $F^r(x) = F(x + \delta_r)$. Since F is nondegenerate, there is an x such that $0 < F(x) < 1$, and it follows by (14.20) that d_k is decreasing in k, so that δ_r is strictly decreasing in r.

---

*This topic may be omitted.

For positive real t let $\varphi(t) = \inf_{0 < r < t} \delta_r$ (r rational in the infimum). Then $\varphi(t)$ is decreasing in t, and

$$(14.21) \qquad\qquad F^t(x) = F(x + \varphi(t))$$

for all x and all positive t. Further, (14.19) implies that $\varphi(st) = \varphi(s) + \varphi(t)$, so that by the theorem on Cauchy's equation [A20] applied to $\varphi(e^x)$, $\varphi(t) = -\beta \log t$, where $\beta > 0$ because $\varphi(t)$ is strictly decreasing. Now (14.21) with $t = e^{x/\beta}$ gives $F(x) = \exp\{e^{-x/\beta} \log F(0)\}$, and so F must be of the same type as

$$(14.22) \qquad\qquad F_1(x) = e^{-e^{-x}}.$$

Example 14.1 shows that this distribution function can arise as a limit of distributions of maxima—that is, F_1 is indeed extremal.

CASE 2. Suppose that $c_{k_0} \neq 1$ for some k_0, which necessarily exceeds 1. Then there exists an x' such that $c_{k_0} x' + d_{k_0} = x'$; but (14.18) then gives $F^{k_0}(x') = F(x')$, so that $F(x')$ is 0 or 1. (In Case 1, F has the type (14.22) and so never assumes the values 0 and 1.)

Now suppose further that, in fact, $F(x') = 0$. Let x_0 be the supremum of those x for which $F(x) = 0$. By passing to a new F of the same type one can arrange that $x_0 = 0$; then $F(x) = 0$ for $x < 0$ and $F(x) > 0$ for $x > 0$. The new F will satisfy (14.18), but with new constants d_k.

If a (new) d_k is distinct from 0, then there is an x near 0 for which the arguments on the two sides of (14.18) have opposite signs. Therefore, $d_k = 0$ for all k, and

$$(14.23) \qquad F^k(x) = F(c_k x), \qquad F^{1/k}(x) = F\left(\frac{x}{c_k}\right)$$

for all k and x. This implies that $F^{j/k}(x) = F(x c_j / c_k)$. For positive rational $r = j/k$, put $\gamma_r = c_j / c_k$. The definition is again consistent by (14.19), and $F^r(x) = F(\gamma_r x)$. Since $0 < F(x) < 1$ for some x, necessarily positive, it follows by (14.23) that c_k is decreasing in k, so that γ_r is strictly decreasing in r. Put $\psi(t) = \inf_{0 < r < t} \gamma_r$ for positive real t. From (14.19) follows $\psi(st) = \psi(s)\psi(t)$, and by the corollary to the theorem on Cauchy's equation [A20] applied to $\psi(e^x)$, it follows that $\psi(t) = t^{-\xi}$ for some $\xi > 0$. Since $F^t(x) = F(\psi(t)x)$ for all x and for t positive, $F(x) = \exp\{x^{-1/\xi} \log F(1)\}$ for $x > 0$. Thus (take $\alpha = 1/\xi$) F is of the same type as

$$(14.24) \qquad\qquad F_{2,\alpha}(x) = \begin{cases} 0 & \text{if } x < 0, \\ e^{-x^{-\alpha}} & \text{if } x \geq 0. \end{cases}$$

Example 14.2 shows that this case can arise.

CASE 3. Suppose as in Case 2 that $c_{k_0} \neq 1$ for some k_0, so that $F(x')$ is 0 or 1 for some x', but this time suppose that $F(x') = 1$. Let x_1 be the infimum of those x for which $F(x) = 1$. By passing to a new F of the same type, arrange that $x_1 = 0$; then $F(x) < 1$ for $x < 0$ and $F(x) = 1$ for $x \geq 0$. If $d_k \neq 0$, then for some x near 0, one side of (14.18) is 1 and the other is not. Thus $d_k = 0$ for all k, and (14.23) again holds. And

again $\gamma_{j/k} = c_j/c_k$ consistently defines a function satisfying $F^r(x) = F(\gamma_r x)$. Since F is nondegenerate, $0 < F(x) < 1$ for some x, but this time x is necessarily negative, so that c_k is increasing.

The same analysis as before shows that there is a positive ξ such that $F^t(x) = F(t^\xi x)$ for all x and for t positive. Thus $F(x) = \exp\{(-x)^{1/\xi} \log F(-1)\}$ for $x < 0$, and F is of the type

$$(14.25) \qquad F_{3,\alpha}(x) = \begin{cases} e^{-(-x)^\alpha} & \text{if } x \le 0, \\ 1 & \text{if } x \ge 0. \end{cases}$$

Example 14 3 shows that this distribution function is indeed extremal.

This completely characterizes the class of extremal distributions:

Theorem 14.3. *The class of extremal distribution functions consists exactly of the distribution functions of the types* (14.22), (14.24), *and* (14.25).

It is possible to go on and characterize the *domains of attraction*. That is. it is possible for each extremal distribution function F to describe the class of G satisfying (14.17) for some constants a_n and b_n—the class of G *attracted* to F.[†]

PROBLEMS

14.1. The general nondecreasing function F has at most countably many discontinuities. Prove this by considering the open intervals

$$\left(\sup_{u < x} F(u), \inf_{\iota > x} F(\upsilon) \right)$$

—each nonempty one contains a rational.

14.2. For distribution functions F, the second proof of Theorem 14.1 shows how to construct a measure μ on $(R^1, \mathcal{R}^1)$ such that $\mu(a, b] = F(b) - F(a)$.

(a) Extend to the case of bounded F.

(b) Extend to the general case. *Hint:* Let $F_n(x)$ be $-n$ or $F(x)$ or n as $F(x) < -n$ or $-n \le F(x) < n$ or $n \le F(x)$. Construct the corresponding μ_n and define $\mu(A) = \lim_n \mu_n(A)$.

14.3. (a) Suppose that X has a continuous, strictly increasing distribution function F. Show that the random variable $F(X)$ is uniformly distributed over the unit interval in the sense that $P[F(X) \le u] = u$ for $0 \le u \le 1$. Passing from X to $F(X)$ is called the *probability transformation*.

(b) Show that the function $\varphi(u)$ defined by (14.5) satisfies $F(\varphi(u) -) \le u \le F(\varphi(u))$ and that, if F is continuous (but not necessarily strictly increasing), then $F(\varphi(u)) = u$ for $0 < u < 1$.

(c) Show that $P[F(X) < u] = F(\varphi(u) -)$ and hence that the result in part (a) holds as long as F is continuous.

[†]This theory is associated with the names of Fisher, Fréchet, Gnedenko, and Tippet. For further information, see GALAMBOS.

14.4. ↑   Let C be the set of continuity points of F.

(a) Show that for every Borel set A, $P[F(X) \in A, X \in C]$ is at most the Lebesgue measure of A.

(b) Show that if F is continuous at each point of $F^{-1}A$, then $P[F(X) \in A]$ is at most the Lebesgue measure of A.

14.5. The *Lévy distance* $d(F, G)$ between two distribution functions is the infimum of those ϵ such that $G(x - \epsilon) - \epsilon \le F(x) \le G(x + \epsilon) + \epsilon$ for all x. Verify that this is a metric on the set of distribution functions. Show that a necessary and sufficient condition for $F_n \Rightarrow F$ is that $d(F_n, F) \to 0$.

14.6. 12.3↑   A Borel function satisfying Cauchy's equation [A20] is automatically bounded in some interval and hence satisfies $f(x) = xf(1)$. *Hint:* Take K large enough that $\lambda[x: x > s, |f(x)| \le K] > 0$. Apply Problem 12.3 and conclude that f is bounded in some interval to the right of 0

14.7. ↑   Consider sets S of reals that are linearly independent over the field of rationals in the sense that $n_1 x_1 + \cdots + n_k x_k = 0$ for distinct points x_i in S and integers n_i (positive or negative) is impossible unless $n_i \equiv 0$.

(a) By Zorn's lemma find a maximal such S. Show that it is a *Hamel basis*. That is, show that each real x can be written uniquely as $x = n_1 x_1 + \cdots + n_k x_k$ for distinct points x_i in S and integers n_i.

(b) Define f arbitrarily on S, and define it elsewhere by $f(n_1 x_1 + \cdots + n_k x_k) = n_1 f(x_1) + \cdots + n_k f(x_k)$. Show that f satisfies Cauchy's equation but need not satisfy $f(x) = xf(1)$.

(c) By means of Problem 14.6 give a new construction of a nonmeasurable function and a nonmeasurable set.

14.8. 14.5↑   (a) Show that if a distribution function F is everywhere continuous, then it is uniformly continuous.

(b) Let $\delta_F(\epsilon) = \sup[F(x) - F(y): |x - y| \le \epsilon]$ be the modulus of continuity of F. Show that $d(F, G) < \epsilon$ implies that $\sup_x |F(x) - G(x)| \le \epsilon + \delta_F(\epsilon)$.

(c) Show that, if $F_n \Rightarrow F$ and F is everywhere continuous, then $F_n(x) \to F(x)$ uniformly in x. What if F is continuous over a closed interval?

14.9. Show that (14.24) and (14.25) are everywhere infinitely differentiable, although not analytic.

CHAPTER 3

Integration

SECTION 15. THE INTEGRAL

Expected values of simple random variables and Riemann integrals of continuous functions can be brought together with other related concepts under a general theory of integration, and this theory is the subject of the present chapter.

Definition

Throughout this section, f, g, and so on will denote real measurable functions, the values $\pm\infty$ allowed, on a measure space $(\Omega, \mathscr{F}, \mu)$.[†] The object is to define and study the definite integral

$$\int f d\mu = \int_\Omega f(\omega)\, d\mu(\omega) = \int_\Omega f(\omega)\mu(d\omega).$$

Suppose first that f is nonnegative. For each finite decomposition $\{A_i\}$ of Ω into $\mathscr{F}$-sets, consider the sum

(15.1) $$\sum_i \left[\inf_{\omega \in A_i} f(\omega)\right] \mu(A_i).$$

In computing the products here, the conventions about infinity are

(15.2)
$$0 \cdot \infty = \infty \cdot 0 = 0,$$
$$x \cdot \infty = \infty \cdot x = \infty \qquad \text{if } 0 < x < \infty,$$
$$\infty \cdot \infty = \infty.$$

[†]Although the definitions (15.3) and (15.6) apply even if f is not measurable $\mathscr{F}$, the proofs of most theorems about integration do use the assumption of measurability in one way or another. For the role of measurability, and for alternative definitions of the integral, see the problems.

199

The reasons for these conventions will become clear later. Also in force are the conventions of Section 10 for sums and limits involving infinity; see (10.3) and (10.4). If A_i is empty, the infimum in (15.1) is by the standard convention ∞; but then $\mu(A_i) = 0$, so that by the convention (15.2), this term makes no contribution to the sum (15.1).

The integral of f is defined as the supremum of the sums (15.1):

$$(15.3) \qquad \int f d\mu = \sup_i \sum \left[\inf_{\omega \in A_i} f(\omega) \right] \mu(A_i).$$

The supremum here extends over all finite decompositions $\{A_i\}$ of Ω into $\mathscr{F}$-sets.

For general f, consider its *positive part*,

$$(15.4) \qquad f^+(\omega) = \begin{cases} f(\omega) & \text{if } 0 \le f(\omega) \le \infty, \\ 0 & \text{if } -\infty \le f(\omega) \le 0 \end{cases}$$

and its *negative part*,

$$(15.5) \qquad f^-(\omega) = \begin{cases} -f(\omega) & \text{if } -\infty \le f(\omega) \le 0, \\ 0 & \text{if } 0 \le f(\omega) \le \infty. \end{cases}$$

These functions are nonnegative and measurable, and $f = f^+ - f^-$. The general integral is defined by

$$(15.6) \qquad \int f d\mu = \int f^+ d\mu - \int f^- d\mu,$$

unless $\int f^+ d\mu = \int f^- d\mu = \infty$, in which case f has no integral.

If $\int f^+ d\mu$ and $\int f^- d\mu$ are both finite, then f is *integrable*, or integrable μ, or summable, and has (15.6) as its *definite integral*. If $\int f^+ d\mu = \infty$ and $\int f^- d\mu < \infty$, then f is not integrable but in accordance with (15.6) is assigned ∞ as its definite integral. Similarly, if $\int f^+ d\mu < \infty$ and $\int f^- d\mu = \infty$, then f is not integrable but has definite integral $-\infty$. Note that f can have a definite integral without being integrable; it fails to have a definite integral if and only if its positive and negative parts both have infinite integrals.

The really important case of (15.6) is that in which $\int f^+ d\mu$ and $\int f^- d\mu$ are both finite. Allowing infinite integrals is a convention that simplifies the statements of various theorems, especially theorems involving nonnegative functions. Note that (15.6) is defined unless it involves "$\infty - \infty$"; if one term on the right is ∞ and the other is a finite real x, the difference is defined by the conventions $\infty - x = \infty$ and $x - \infty = -\infty$.

The extension of the integral from the nonnegative case to the general case is consistent: (15.6) agrees with (15.3) if f is nonnegative, because then $f^- \equiv 0$.

Nonnegative Functions

It is convenient first to analyze nonnegative functions.

Theorem 15.1. (i) *If* $f = \sum_i x_i I_{A_i}$ *is a nonnegative simple function,* $\{A_i\}$ *being a finite decomposition of* Ω *into* $\mathscr{F}$-*sets, then* $\int f d\mu = \sum_i x_i \mu(A_i)$.

(ii) *If* $0 \le f(\omega) \le g(\omega)$ *for all* ω, *then* $\int f d\mu \le \int g d\mu$.

(iii) *If* $0 \le f_n(\omega) \uparrow f(\omega)$ *for all* ω, *then* $0 \le \int f_n d\mu \uparrow \int f d\mu$.

(iv) *For nonnegative functions* f *and* g *and nonnegative constants* α *and* β,
$\int (\alpha f + \beta g) d\mu = \alpha \int f d\mu + \beta \int g d\mu$.

In part (iii) the essential point is that $\int f d\mu = \lim_n \int f_n d\mu$, and it is important to understand that both sides of this equation may be ∞. If $f_n = I_{A_n}$ and $f = I_A$, where $A_n \uparrow A$, the conclusion is that μ is continuous from below (Theorem 10.2(i)): $\lim_n \mu(A_n) = \mu(A)$; this equation often takes the form $\infty = \infty$.

PROOF OF (i). Let $\{B_j\}$ be a finite decomposition of Ω and let β_j be the infimum of f over B_j. If $A_i \cap B_j \ne \varnothing$, then $\beta_j \le x_i$; therefore, $\sum_j \beta_j \mu(B_j) = \sum_{ij} \beta_j \mu((A_i \cap B_j)) \le \sum_{ij} x_i \mu(A_i \cap B_j) = \sum_i x_i \mu(A_i)$. On the other hand, there is equality here if $\{B_j\}$ coincides with $\{A_i\}$. ∎

PROOF OF (ii). The sums (15.1) obviously do not decrease if f is replaced by g. ∎

PROOF OF (iii). By (ii) the sequence $\int f_n d\mu$ is nondecreasing and bounded above by $\int f d\mu$. It therefore suffices to show that $\int f d\mu \le \lim_n \int f_n d\mu$, or that

$$(15.7) \qquad \lim_n \int f_n d\mu \ge S = \sum_{i=1}^{m} v_i \mu(A_i)$$

if $A_1, \ldots, A_m$ is any decomposition of Ω into $\mathscr{F}$-sets and $v_i = \inf_{\omega \in A_i} f(\omega)$.

In order to see the essential idea of the proof, which is quite simple, suppose first that S is finite and all the v_i and $\mu(A_i)$ are positive and finite. Fix an ϵ that is positive and less than each v_i, and put $A_{in} = [\omega \in A_i : f_n(\omega) > v_i - \epsilon]$. Since $f_n \uparrow f$, $A_{in} \uparrow A_i$. Decompose Ω into $A_{1n}, \ldots, A_{mn}$ and the complement of their union, and observe that, since μ is continuous from below,

$$(15.8) \qquad \int f_n d\mu \ge \sum_{i=1}^{m} (v_i - \epsilon)\mu(A_{in}) \to \sum_{i=1}^{m} (v_i - \epsilon)\mu(A_i)$$

$$= S - \epsilon \sum_{i=1}^{m} \mu(A_i).$$

Since the $\mu(A_i)$ are all finite, letting $\epsilon \to 0$ gives (15.7).

Now suppose only that S is finite. Each product $v_i\mu(A_i)$ is then finite; suppose it is positive for $i \le m_0$ and 0 for $i > m_0$. (Here $m_0 \le m$; if the product is 0 for all i, then $S = 0$ and (15.7) is trivial.) Now v_i and $\mu(A_i)$ are positive and finite for $i \le m_0$ (one or the other may be ∞ for $i > m_0$). Define A_{in} as before, but only for $i \le m_0$. This time decompose Ω into $A_{1n}, \ldots, A_{m_0 n}$ and the complement of their union. Replace m by m_0 in (15.8) and complete the proof as before.

Finally, suppose that $S = \infty$. Then $v_{i_0}\mu(A_{i_0}) = \infty$ for some i_0, so that v_{i_0} and $\mu(A_{i_0})$ are both positive and at least one is ∞. Suppose $0 < x < v_{i_0} \le \infty$ and $0 < y < \mu(A_{i_0}) \le \infty$, and put $A_{i_0 n} = [\omega \in A_{i_0}: f_n(\omega) > x]$. From $f_n \uparrow f$ follows $A_{i_0 n} \uparrow A_{i_0}$; hence $\mu(A_{i_0 n}) > y$ for n exceeding some n_0. But then (decompose Ω into $A_{i_0 n}$ and its complement) $\int f_n \, d\mu \ge x\mu(A_{i_0 n}) \ge xy$ for $n > n_0$, and therefore $\lim_n \int f_n \, d\mu \ge xy$. If $v_{i_0} = \infty$, let $x \to \infty$, and if $\mu(A_{i_0}) = \infty$, let $y \to \infty$. In either case (15.7) follows: $\lim_n \int f_n \, d\mu = \infty$.   ∎

PROOF OF (iv).   Suppose at first that $f = \sum_i x_i I_{A_i}$ and $g = \sum_j y_j I_{B_j}$ are simple. Then $\alpha f + \beta g = \sum_{ij}(\alpha x_i + \beta y_j)I_{A_i \cap B_j}$, and so

$$\int (\alpha f + \beta g)\, d\mu = \sum_{ij} (\alpha x_i + \beta y_j)\mu(A_i \cap B_j)$$

$$= \alpha \sum_i x_i\mu(A_i) + \beta \sum_j y_j\mu(B_j) = \alpha \int f \, d\mu + \beta \int g \, d\mu.$$

Note that the argument is valid if some of α, β, x_i, y_j are infinite. Apart from this possibility, the ideas are as in the proof of (5.21).

For general nonnegative f and g, there exist by Theorem 13.5 simple functions f_n and g_r such that $0 \le f_n \uparrow f$ and $0 \le g_n \uparrow g$. But then $0 \le \alpha f_n + \beta g_n \uparrow \alpha f + \beta g$ and $\int(\alpha f_n + \beta g_n)\, d\mu = \alpha \int f_n \, d\mu + \beta \int g_n \, d\mu$, so that (iv) follows from (iii).   ∎

By part (i) of Theorem 15.1, the expected values of simple random variables in Chapter 1 are integrals: $E[X] = \int X(\omega)P(d\omega)$. This also covers the step functions in Section 1 (see (1.6)). The relation between the Riemann integral and the integral as defined here will be studied in Section 17.

Example 15.1.   Consider the line $(R^1, \mathscr{R}^1, \lambda)$ with Lebesgue measure. Suppose that $-\infty < a_0 \le a_1 \le \cdots \le a_m < \infty$, and let f be the function with nonnegative value x_i on $(a_{i-1}, a_i]$, $i = 1, \ldots, m$, and value 0 on $(-\infty, a_0]$ and (a_m, ∞). By part (i) of Theorem 15.1, $\int f \, d\lambda = \sum_{i=1}^m x_i(a_i - a_{i-1})$ because of the convention $0 \cdot \infty = 0$—see (15.2). If the "area under the curve" to the left of a_0 and to the right of a_m is to be 0, this convention is inevitable. From $\infty \cdot 0 = 0$ it follows that $\int f \, d\lambda = 0$ if f is ∞ at a single point (say) and 0 elsewhere.

If $f = I_{(a,\infty)}$, the area-under-the-curve point of view makes $\int f\,d\mu = \infty$ natural. Hence the second convention in (15.2), which also requires that the integral be infinite if f is ∞ on a nonempty interval and 0 elsewhere. ■

Recall that *almost everywhere* means outside a set of measure 0.

Theorem 15.2. *Suppose that f and g are nonnegative.*

(i) *If $f = 0$ almost everywhere, then $\int f\,d\mu = 0$.*

(ii) *If $\mu[\omega: f(\omega) > 0] > 0$, then $\int f\,d\mu > 0$.*

(iii) *If $\int f\,d\mu < \infty$, then $f < \infty$ almost everywhere.*

(iv) *If $f \le g$ almost everywhere, then $\int f\,d\mu \le \int g\,d\mu$.*

(v) *If $f = g$ almost everywhere, then $\int f\,d\mu = \int g\,d\mu$.*

PROOF. Suppose that $f = 0$ almost everywhere. If A_i meets $[\omega: f(\omega) = 0]$, then the infimum in (15.1) is 0; otherwise, $\mu(A_i) = 0$. Hence each sum (15.1) is 0, and (i) follows.

If $A_\epsilon = [\omega: f(\omega) \ge \epsilon]$, then $A_\epsilon \uparrow [\omega: f(\omega) > 0]$ as $\epsilon \downarrow 0$, so that under the hypothesis of (ii) there is a positive ϵ for which $\mu(A_\epsilon) > 0$. Decomposing Ω into A_ϵ and its complement shows that $\int f\,d\mu \ge \epsilon\mu(A_\epsilon) > 0$.

If $\mu[f = \infty] > 0$, decompose Ω into $[f = \infty]$ and its complement: $\int f\,d\mu \ge \infty \cdot \mu[f = \infty] = \infty$ by the conventions. Hence (iii).

To prove (iv), let $G = [f \le g]$. For any finite decomposition $\{A_1, \ldots, A_m\}$ of Ω,

$$\sum \left[\inf_{A_i} f\right]\mu(A_i) = \sum \left[\inf_{A_i} f\right]\mu(A_i \cap G) \le \sum \left[\inf_{A_i \cap G} f\right]\mu(A_i \cap G)$$

$$\le \sum \left[\inf_{A_i \cap G} g\right]\mu(A_i \cap G) \le \int g\,d\mu,$$

where the last inequality comes from a consideration of the decomposition $A_1 \cap G, \ldots, A_m \cap G, G^c$. This proves (iv), and (v) follows immediately. ■

Suppose that $f = g$ almost everywhere, where f and g need not be nonnegative. If f has a definite integral, then since $f^+ = g^+$ and $f^- = g^-$ almost everywhere, it follows by Theorem 15.2(v) that g also has a definite integral and $\int f\,d\mu = \int g\,d\mu$.

Uniqueness

Although there are various ways to frame the definition of the integral, they are all equivalent—they all assign the same value to $\int f\,d\mu$. This is because the integral is uniquely determined by certain simple properties it is natural to require of it.

It is natural to want the integral to have properties (i) and (iii) of Theorem 15.1. But these uniquely determine the integral for nonnegative functions: For f nonnegative, there exist by Theorem 13.5 simple functions f_n such that $0 \le f_n \uparrow f$; by (iii), $\int f\,d\mu$ must be $\lim_n \int f_n\,d\mu$, and (i) determines the value of each $\int f_n\,d\mu$.

Property (i) can itself be derived from (iv) (linearity) together with the assumption that $\int I_A \, d\mu = \mu(A)$ for indicators I_A: $\int (\sum_i x_i I_{A_i}) \, d\mu = \sum_i x_i \int I_{A_i} \, d\mu = \sum_i x_i \mu(A_i)$.

If (iv) of Theorem 15.1 is to persist when the integral is extended beyond the class of nonnegative functions, $\int f d\mu$ must be $\int (f^+ - f^-) \, d\mu = \int f^+ \, d\mu - \int f^- \, d\mu$, which makes the definition (15.6) inevitable.

PROBLEMS

These problems outline alternative definitions of the integral and clarify the role measurability plays. Call (15.3) the *lower integral*, and write it as

$$(15\,9) \qquad \int_* f d\mu = \sup \sum_i \left[\inf_{\omega \in A_i} f(\omega) \right] \mu(A_i)$$

to distinguish it from the *upper integral*

$$(15.10) \qquad \int^* f d\mu = \inf \sum_i \left[\sup_{\omega \in A_i} f(\omega) \right] \mu(A_i).$$

The infimum in (15 10), like the supremum in (15.9), extends over all finite partitions $\{A_i\}$ of Ω into $\mathscr{F}$-sets.

15.1. Suppose that f is measurable and nonnegative. Show that $\int^* f d\mu = \infty$ if $\mu[\omega: f(\omega) > 0] = \infty$ or if $\mu[\omega: f(\omega) > a] > 0$ for all a

There are many functions familiar from calculus that ought to be integrable but are of the types in the preceding problem and hence have infinite upper integral. Examples are $x^{-2} I_{(1,\infty)}(x)$ and $x^{-1/2} I_{(0,1)}(x)$. Therefore, (15.10) is inappropriate as a definition of $\int f d\mu$ for nonnegative f. The only problem with (15.10), however, is that it treats infinity the wrong way. To see this, and to focus on essentials, assume that $\mu(\Omega) < \infty$ and that f is bounded, although not necessarily nonnegative or measurable $\mathscr{F}$.

15.2. ↑ (a) Show that

$$\sum_i \left[\inf_{\omega \in A_i} f(\omega) \right] \mu(A_i) \le \sum_j \left[\inf_{\omega \in B_j} f(\omega) \right] \mu(B_j)$$

if $\{B_j\}$ refines $\{A_i\}$. Prove a dual relation for the sums in (15.10) and conclude that

$$(15.11) \qquad \int_* f d\mu \le \int^* f d\mu.$$

(b) Now assume that f is measurable $\mathscr{F}$ and let M be a bound for $|f|$. Consider the partition $A_i = [\omega: i\epsilon < f(\omega) \le (i+1)\epsilon]$, where i ranges from $-N$

to N and N is large enough that $N\epsilon > M$. Show that

$$\sum_i \left[\sup_{\omega \in A_i} f(\omega) \right] \mu(A_i) - \sum_i \left[\inf_{\omega \in A_i} f(\omega) \right] \mu(A_i) \le \epsilon \mu(\Omega).$$

Conclude that

(15.12) $$\int_* f\, d\mu = \int^* f\, d\mu.$$

To define the integral as the common value in (15.12) is the *Darboux-Young* approach. The advantage of (15.3) as a definition is that (in the nonnegative case) it applies at once to unbounded f and infinite μ

15.3. 3.2 15.2↑   For $A \subset \Omega$, define $\mu^*(A)$ and $\mu_*(A)$ by (3.9) and (3.10) with μ in place of P. Show that $\int^* I_A \, d\mu = \mu^*(A)$ and $\int_* I_A \, d\mu = \mu_*(A)$ for every A. Therefore, (15.12) can fail if f is not measurable $\mathscr{F}$. (Where was measurability used in the proof of (15.12)?)

The definitions (15.3) and (15.6) always make formal sense (for finite $\mu(\Omega)$ and $\sup|f|$), but they are reasonable—accord with intuition—only if (15.12) holds. Under what conditions *does* it hold?

15.4. 10.5 15.3↑   (a) Suppose of f that *there exist an $\mathscr{F}$-set A and a function g, measurable $\mathscr{F}$, such that* $\mu(A) = 0$ *and* $[f \ne g] \subset A$. This is the same thing as assuming that $\mu^*[f \ne g] = 0$, or assuming that f is measurable with respect to $\mathscr{F}$ completed with respect to μ. Show that (15.12) holds.

(b) Show that if (15.12) holds, then so does the italicized condition in part (a).

Rather than assume that f is measurable $\mathscr{F}$, one can assume that it satisfies the italicized condition in Problem 15.4(a)—which in case $(\Omega, \mathscr{F}, \mu)$ is complete is the same thing anyway. For the next three problems, assume that $\mu(\Omega) < \infty$ and that f is measurable $\mathscr{F}$ and bounded.

15.5. ↑   Show that for positive ϵ there exists a finite partition $\{A_i\}$ such that, if $\{B_j\}$ is any finer partition and $\omega_j \in B_j$, then

$$\left| \int f\, d\mu - \sum_j f(\omega_j)\mu(B_j) \right| < \epsilon.$$

15.6. ↑   Show that

$$\int f\, d\mu = \lim_n \sum_{|k| \le n2^n} \frac{k-1}{2^n} \mu\left[\omega : \frac{k-1}{2^n} \le f(\omega) < \frac{k}{2^n} \right].$$

The limit on the right here is *Lebesgue's* definition of the integral.

15.7. ↑ Suppose that the integral is *defined* for simple nonnegative functions by $\int (\Sigma_i x_i I_{A_i}) d\mu = \Sigma_i x_i \mu(A_i)$. Suppose that f_n and g_n are simple and nondecreasing and have a common limit: $0 \le f_n \uparrow f$ and $0 \le g_n \uparrow f$. Adapt the arguments used to prove Theorem 15.1(iii) and show that $\lim_n \int f_n d\mu = \lim_n \int g_n d\mu$. Thus, in the nonnegative case, $\int f d\mu$ can (Theorem 13.5) consistently be *defined* as $\lim_n \int f_n d\mu$ for simple functions for which $0 \le f_n \uparrow f$.

SECTION 16. PROPERTIES OF THE INTEGRAL

Equalities and Inequalities

By definition, the requirement for integrability of f is that $\int f^+ d\mu$ and $\int f^- d\mu$ both be finite, which is the same as the requirement that $\int f^+ d\mu + \int f^- d\mu < \infty$ and hence is the same as the requirement that $\int (f^+ + f^-) d\mu < \infty$ (Theorem 15.1(iv)). Since $f^+ + f^- = |f|$, f is integrable if and only if

$$(16.1) \qquad \int |f| d\mu < \infty.$$

It follows that if $|f| \le |g|$ almost everywhere and g is integrable, then f is integrable as well. If $\mu(\Omega) < \infty$, a bounded f is integrable.

Theorem 16.1. (i) *Monotonicity*: *If f and g are integrable and $f \le g$ almost everywhere, then*

$$(16.2) \qquad \int f d\mu \le \int g d\mu.$$

(ii) *Linearity*: *If f and g are integrable and α, β are finite real numbers, then $\alpha f + \beta g$ is integrable and*

$$(16.3) \qquad \int (\alpha f + \beta g) d\mu = \alpha \int f d\mu + \beta \int g d\mu.$$

PROOF OF (i). For nonnegative f and g such that $f \le g$ almost everywhere, (16.2) follows by Theorem 15.2(iv). And for general integrable f and g, if $f \le g$ almost everywhere, then $f^+ \le g^+$ and $f^- \ge g^-$ almost everywhere, and so (16.2) follows by the definition (15.6). ∎

PROOF OF (ii). First, $\alpha f + \beta g$ is integrable because, by Theorem 15.1,

$$\int |\alpha f + \beta g| d\mu \le \int (|\alpha| \cdot |f| + |\beta| \cdot |g|) d\mu$$

$$= |\alpha| \int |f| d\mu + |\beta| \int |g| d\mu < \infty.$$

By Theorem 15.1(iv) and the definition (15.6), $\int (\alpha f)\,d\mu = \alpha \int f\,d\mu$—consider separately the cases $\alpha \geq 0$ and $\alpha < 0$. Therefore, it is enough to check (16.3) for the case $\alpha = \beta = 1$. By definition, $(f + g)^+ - (f + g)^- = f + g = f^+ - f^- + g^+ - g^-$ and therefore $(f + g)^+ + f^- + g^- = (f + g)^- + f^+ + g^+$. All these functions being nonnegative, $\int (f + g)^+ \,d\mu + \int f^- \,d\mu + \int g^- \,d\mu = \int (f + g)^- \,d\mu + \int f^+ \,d\mu + \int g^+ \,d\mu$, which can be rearranged to give $\int (f + g)^+ \,d\mu - \int (f + g)^- \,d\mu = \int f^+ \,d\mu - \int f^- \,d\mu + \int g^+ \,d\mu - \int g^- \,d\mu$. But this reduces to (16.3). ∎

Since $-|f| \leq f \leq |f|$, it follows by Theorem 16.1 that

$$(16.4) \qquad \left| \int f\,d\mu \right| \leq \int |f|\,d\mu$$

for integrable f. Applying this to integrable f and g gives

$$(16.5) \qquad \left| \int f\,d\mu - \int g\,d\mu \right| \leq \int |f - g|\,d\mu.$$

Example 16.1. Suppose that Ω is countable, that $\mathscr{F}$ consists of all the subsets of Ω, and that μ is counting measure: each singleton has measure 1. To be definite, take $\Omega = \{1, 2, \ldots\}$. A function is then a sequence $x_1, x_2, \ldots$. If x_{nm} is x_m or 0 as $m \leq n$ or $m > n$, the function corresponding to $x_{n1}, x_{n2}, \ldots$ has integral $\sum_{m=1}^{n} x_m$ by Theorem 15.1(i) (consider the decomposition $\{1\}, \ldots, \{n\}, \{n + 1, n + 2, \ldots\}$). It follows by Theorem 15.1(iii) that in the nonnegative case the integral of the function given by $\{x_m\}$ is the sum $\sum_m x_m$ (finite or infinite) of the corresponding infinite series. In the general case the function is integrable if and only if $\sum_{m=1}^{\infty} |x_m|$ is a convergent infinite series, in which case the integral is $\sum_{m=1}^{\infty} x_m^+ - \sum_{m=1}^{\infty} x_m^-$.

The function $x_m = (-1)^{m+1} m^{-1}$ is not integrable by this definition and even fails to have a definite integral, since $\sum_{m=1}^{\infty} x_m^+ = \sum_{m=1}^{\infty} x_m^- = \infty$. This invites comparison with the ordinary theory of infinite series, according to which the alternating harmonic series does converge in the sense that $\lim_M \sum_{m=1}^{M} (-1)^{m+1} m^{-1} = \log 2$. But since this says that the sum of the *first* M terms has a limit, it requires that the elements of the space Ω be ordered. If Ω consists not of the positive integers but, say, of the integer lattice points in 3-space, it has no canonical linear ordering. And if $\sum_m x_m$ is to have the same finite value no matter what the order of summation, the series must be absolutely convergent.[†] This helps to explain why f is defined to be integrable only if $\int f^+ \,d\mu$ and $\int f^- \,d\mu$ are both finite. ∎

Example 16.2. In connection with Example 15.1, consider the function $f = 3 I_{(a, \infty)} - 2 I_{(-\infty, a)}$. There is no natural value for $\int f\,d\lambda$ (it is "$\infty - \infty$"), and none is assigned by the definition.

[†] RUDIN$_1$, p. 76.

If a function f is bounded on bounded intervals, then each function $f_n = fI_{(-n,n)}$ is integrable with respect to λ. Since $f = \lim_n f_n$, the limit of $\int f_n \, d\lambda$, if it exists, is sometimes called the "principal value" of the integral of f. Although it is natural for some purposes to integrate symmetrically about the origin, this is not the right definition of the integral in the context of general measure theory. The functions $g_n = fI_{(-n,n+1)}$ for example also converge to f, and $\int g_n \, d\lambda$ may have some other limit, or none at all; $f(x) = x$ is a case in point. There is no general reason why f_n should take precedence over g_n.

As in the preceding example, $f = \sum_{k=1}^{\infty} (-1)^k k^{-1} I_{(k,k+1)}$ has no integral, even though the $\int f_n \, d\lambda$ above converge. ∎

Integration to the Limit

The first result, the *monotone convergence theorem*, essentially restates Theorem 15.1(iii).

Theorem 16.2. *If* $0 \leq f_n \uparrow f$ *almost everywhere, then* $\int f_n \, d\mu \uparrow \int f d\mu$.

PROOF. If $0 \leq f_n \uparrow f$ on a set A with $\mu(A^c) = 0$, then $0 \leq f_n I_A \uparrow fI_A$ holds everywhere, and it follows by Theorem 15.1(iii) and the remark following Theorem 15.2 that $\int f_n \, d\mu = \int f_n I_A \, d\mu \uparrow \int fI_A \, d\mu = \int f d\mu$. ∎

As the functions in Theorem 16.2 are nonnegative almost everywhere, all the integrals exist. The conclusion of the theorem is that $\lim_n \int f_n \, d\mu$ and $\int f d\mu$ are both infinite or both finite and in the latter case are equal.

Example 16.3. Consider the space $\{1, 2, \ldots\}$ together with counting measure, as in Example 16.1. If for each m one has $0 \leq x_{nm} \uparrow x_m$ as $n \to \infty$, then $\lim_n \sum_m x_{nm} = \sum_m x_m$, a standard result about infinite series. ∎

Example 16.4. If μ is a measure on $\mathscr{F}$, and $\mathscr{F}_0$ is a σ-field contained in $\mathscr{F}$, then the restriction μ_0 of μ to $\mathscr{F}_0$ is another measure (Example 10.4). If $f = I_A$ and $A \in \mathscr{F}_0$, then

$$\int f d\mu = \int f d\mu_0,$$

the common value being $\mu(A) = \mu_0(A)$. The same is true by linearity for nonnegative simple functions measurable $\mathscr{F}_0$. It holds by Theorem 16.2 for all nonnegative f that are measurable $\mathscr{F}_0$ because (Theorem 13.5) $0 \leq f_n \uparrow f$ for simple functions f_n that are measurable $\mathscr{F}_0$. For functions measurable $\mathscr{F}_0$, integration with respect to μ is thus the same thing as integration with respect to μ_0. ∎

In this example a property was extended by linearity from indicators to nonnegative simple functions and thence to the general nonnegative function by a monotone passage to the limit. This is a technique of very frequent application.

Example 16.5. For a finite or infinite sequence of measures μ_n on $\mathscr{F}$, $\mu(A) = \sum_n \mu_n(A)$ defines another measure (countably additive because [A27] sums can be reversed in a nonnegative double series). For indicators f,

$$\int f\, d\mu = \sum_n \int f\, d\mu_n,$$

and by linearity the same holds for simple $f \ge 0$. If $0 \le f_k \uparrow f$ for simple f_k, then by Theorem 16.2 and Example 16.3, $\int f\, d\mu = \lim_k \int f_k\, d\mu = \lim_k \sum_n \int f_k\, d\mu_n = \sum_n \lim_k \int f_k\, d\mu_n = \sum_n \int f\, d\mu_n$. The relation in question thus holds for all nonnegative f. ∎

An important consequence of the monotone convergence theorem is *Fatou's lemma*:

Theorem 16.3. *For nonnegative* f_n,

$$(16.6) \qquad \int \liminf_n f_n\, d\mu \le \liminf_n \int f_n\, d\mu.$$

PROOF. If $g_n = \inf_{k \ge n} f_k$, then $0 \le g_n \uparrow g = \liminf_n f_n$, and the preceding two theorems give $\int f_n\, d\mu \ge \int g_n\, d\mu \to \int g\, d\mu$. ∎

Example 16.6. On $(R^1, \mathscr{R}^1, \lambda)$, the functions $f_n = n^2 I_{(0,\, n^{-1})}$ and $f \equiv 0$ satisfy $f_n(x) \to f(x)$ for each x, but $\int f\, d\lambda = 0$ and $\int f_n\, d\lambda = n \to \infty$. This shows that the inequality in (16.6) can be strict and that it is not always possible to integrate to the limit. This phenomenon has been encountered before; see Examples 5.7 and 7.7. ∎

Fatou's lemma leads to *Lebesgue's dominated convergence theorem*:

Theorem 16.4. *If* $|f_n| \le g$ *almost everywhere, where* g *is integrable, and if* $f_n \to f$ *almost everywhere, then* f *and the* f_n *are integrable and* $\int f_n\, d\mu \to \int f\, d\mu$.

PROOF. Assume at the outset, not that the f_n converge, but only that they are dominated by an integrable g, which implies that all the f_n as well

as $f^* = \limsup_n f_n$ and $f_* = \liminf f_n$ are integrable. Since $g + f_n$ and $g - f_n$ are nonnegative, Fatou's lemma gives

$$\int g \, d\mu + \int f_* \, d\mu = \int \liminf_n (g + f_n) \, d\mu$$

$$\leq \liminf_n \int (g + f_n) \, d\mu = \int g \, d\mu + \liminf_n \int f_n \, d\mu,$$

and

$$\int g \, d\mu - \int f^* \, d\mu = \int \liminf_n (g - f_n) \, d\mu$$

$$\leq \liminf_n \int (g - f_r) \, d\mu = \int g \, d\mu - \limsup_n \int f_n \, d\mu.$$

Therefore

$$(16.7) \qquad \int \liminf_n f_n \, d\mu \leq \liminf_n \int f_n \, d\mu$$

$$\leq \limsup_n \int f_n \, d\mu \leq \int \limsup_n f_n \, d\mu.$$

(Compare this with (4.9).)

Now use the assumption that $f_n \to f$ almost everywhere: f is dominated by g and hence is integrable, and the extreme terms in (16.7) agree with $\int f \, d\mu$. ∎

Example 16.6 shows that this theorem can fail if no dominating g exists.

Example 16.7. The Weierstrass M-test for series. Consider the space $\{1, 2, \ldots\}$ together with counting measure, as in Example 16.1. If $|x_{nm}| \leq M_m$ and $\sum_m M_m < \infty$, and if $\lim_n x_{nm} = x_m$ for each m, then $\lim_n \sum_m x_{nm} = \sum_m x_m$. This follows by an application of Theorem 16.4 with the function given by the sequence $M_1, M_2, \ldots$ in the role of g. This is another standard result on infinite series [A28]. ∎

The next result, the *bounded convergence theorem*, is a special case of Theorem 16.4. It contains Theorem 5.4 as a further special case.

Theorem 16.5. *If $\mu(\Omega) < \infty$ and the f_n are uniformly bounded, then $f_n \to f$ almost everywhere implies $\int f_n \, d\mu \to \int f \, d\mu$.*

The next two theorems are simply the series versions of the monotone and dominated convergence theorems.

Theorem 16.6. *If $f_n \geq 0$, then $\int \Sigma_n f_n \, d\mu = \Sigma_n \int f_n \, d\mu$.*

The members of this last equation are both equal either to ∞ or to the same finite, nonnegative real number.

Theorem 16.7. *If $\Sigma_n f_n$ converges almost everywhere and $|\Sigma_{k=1}^n f_k| \leq g$ almost everywhere, where g is integrable, then $\Sigma_n f_n$ and the f_n are integrable and $\int \Sigma_n f_n \, d\mu = \Sigma_n \int f_n \, d\mu$.*

Corollary. *If $\Sigma_n \int |f_n| \, d\mu < \infty$, then $\Sigma_n f_n$ converges absolutely almost everywhere and is integrable, and $\int \Sigma_n f_n \, d\mu = \Sigma_n \int f_n \, d\mu$.*

PROOF. The function $g = \Sigma_n |f_n|$ is integrable by Theorem 16.6 and is finite almost everywhere by Theorem 15.2(iii). Hence $\Sigma_n |f_n|$ and $\Sigma_n f_n$ converge almost everywhere, and Theorem 16.7 applies. ∎

In place of a sequence $\{f_n\}$ of real measurable functions on $(\Omega, \mathscr{F}, \mu)$, consider a family $[f_t: t > 0]$ indexed by a continuous parameter t. Suppose of a measurable f that

$$\lim_{t \to \infty} f_t(\omega) = f(\omega) \tag{16.8}$$

on a set A, where

$$A \in \mathscr{F}, \qquad \mu(\Omega - A) = 0. \tag{16.9}$$

A technical point arises here, since $\mathscr{F}$ need not contain the ω-set where (16.8) holds:

Example 16.8. Let $\mathscr{F}$ consist of the Borel subsets of $\Omega = [0, 1)$, and let H be a nonmeasurable set—a subset of Ω that does not lie in $\mathscr{F}$ (see the end of Section 3). Define $f_t(\omega) = 1$ if ω equals the fractional part $t - \lfloor t \rfloor$ of t and their common value lies in H^c; define $f_t(\omega) = 0$ otherwise. Each f_t is measurable $\mathscr{F}$, but if $f(\omega) \equiv 0$, then the ω-set where (16.8) holds is exactly H. ∎

Because of such examples, the set A above must be assumed to lie in $\mathscr{F}$. (Because of Theorem 13.4, no such assumption is necessary in the case of sequences.)

Suppose that f and the f_t are integrable. If $I_t = \int f_t \, d\mu$ converges to $I = \int f \, d\mu$ as $t \to \infty$, then certainly $I_{t_n} \to I$ for each sequence $\{t_n\}$ going to infinity. But the converse holds as well: If I_t does not converge to I, then there is a positive ϵ such that $|I_{t_n} - I| > \epsilon$ for a sequence $\{t_n\}$ going to infinity. To the question of whether I_{t_n} converges to I the previous theorems apply.

Suppose that (16.8) and $|f_t(\omega)| \leq g(\omega)$ both hold for $\omega \in A$, where A satisfies (16.9) and g is integrable. By the dominated convergence theorem, f and the f_t must then be integrable and $I_{t_n} \to I$ for each sequence $\{t_n\}$ going to infinity. It follows that $\int f_t \, d\mu \to \int f \, d\mu$. In this result t could go continuously to 0 or to some other value instead of to infinity.

Theorem 16.8. *Suppose that* $f(\omega, t)$ *is a measurable and integrable function of* ω *for each* t *in* (a, b). *Let* $\phi(t) = \int f(\omega, t) \mu(d\omega)$.

(i) *Suppose that for* $\omega \in A$, *where* A *satisfies* (16.9), $f(\omega, t)$ *is continuous in* t *at* t_0; *suppose further that* $|f(\omega, t)| \le g(\omega)$ *for* $\omega \in A$ *and* $|t - t_0| < \delta$, *where* δ *is independent of* ω *and* g *is integrable. Then* $\phi(t)$ *is continuous at* t_0.

(ii) *Suppose that for* $\omega \in A$, *where* A *satisfies* (16.9), $f(\omega, t)$ *has in* (a, b) *a derivative* $f'(\omega, t)$; *suppose further that* $|f'(\omega, t)| \le g(\omega)$ *for* $\omega \in A$ *and* $t \in (a, b)$, *where* g *is integrable. Then* $\phi(t)$ *has derivative* $\int f'(\omega, t) \mu(d\omega)$ *on* (a, b).

PROOF   Part (i) is an immediate consequence of the preceding discussion. To prove part (ii), consider a fixed t. If $\omega \in A$, then by the mean-value theorem,

$$\frac{f(\omega, t + h) - f(\omega, t)}{h} = f'(\omega, s),$$

where s lies between t and $t + h$. The ratio on the left goes[†] to $f'(\omega, t)$ as $h \to 0$ and is by hypothesis dominated by the integrable function $g(\omega)$. Therefore,

$$\frac{\phi(t + h) - \phi(t)}{h} = \int \frac{f(\omega, t + h) - f(\omega, t)}{h} \mu(d\omega) \to \int f'(\omega, t) \mu(d\omega). \qquad \blacksquare$$

The condition involving g in part (ii) can be weakened. It suffices to assume that for each t there is an integrable $g(\omega, t)$ such that $|f'(\omega, s)| \le g(\omega, t)$ for $\omega \in A$ and all s in some neighborhood of t.

Integration over Sets

The integral of f over a set A in $\mathscr{F}$ is defined by

(16.10) $$\int_A f \, d\mu = \int I_A f \, d\mu.$$

The definition applies if f is defined only on A in the first place (set $f = 0$ outside A). Notice that $\int_A f \, d\mu = 0$ if $\mu(A) = 0$.

All the concepts and theorems above carry over in an obvious way to integrals over A. Theorems 16.6 and 16.7 yield this result:

Theorem 16.9. *If* $A_1, A_2, \dots$ *are disjoint, and if* f *is either nonnegative or integrable, then* $\int_{\bigcup_n A_n} f \, d\mu = \sum_n \int_{A_n} f \, d\mu$.

[†]Letting h go to 0 through a sequence shows that each $f'(\cdot, t)$ is measurable $\mathscr{F}$ on A; take it to be 0, say, elsewhere.

The integrals (16.10) usually suffice to determine f:

Theorem 16.10. (i) *If f and g are nonnegative and $\int_A f d\mu = \int_A g d\mu$ for all A in $\mathscr{F}$, and if μ is σ-finite, then $f = g$ almost everywhere.*

(ii) *If f and g are integrable and $\int_A f d\mu = \int_A g d\mu$ for all A in $\mathscr{F}$, then $f = g$ almost everywhere.*

(iii) *If f and g are integrable and $\int_A f d\mu = \int_A g d\mu$ for all A in $\mathscr{P}$, where $\mathscr{P}$ is a π-system generating $\mathscr{F}$ and Ω is a finite or countable union of $\mathscr{P}$-sets, then $f = g$ almost everywhere.*

PROOF. Suppose that f and g are nonnegative and that $\int_A f d\mu \le \int_A g d\mu$ for all A in $\mathscr{F}$. If μ is σ-finite, there are $\mathscr{F}$-sets A_n such that $A_n \uparrow \Omega$ and $\mu(A_n) < \infty$. If $B_n = [0 \le g < f, \ g \le n]$, then the hypothesized inequality applied to $A_n \cap B_n$ implies $\int_{A_n \cap B_n} f d\mu \le \int_{A_n \cap B_n} g d\mu < \infty$ (finite because $A_n \cap B_n$ has finite measure and g is bounded there) and hence $\int I_{A_n \cap B_n}(f - g) d\mu = 0$. But then by Theorem 15.2(ii), the integrand is 0 almost everywhere, and so $\mu(A_n \cap B_n) = 0$. Therefore, $\mu[0 \le g < f, \ g < \infty] = 0$, so that $f \le g$ almost everywhere; (i) follows.

The argument for (ii) is simpler: If f and g are integrable and $\int_A f d\mu \le \int_A g d\mu$ for all A in $\mathscr{F}$, then $\int I_{[g < f]}(f - g) d\mu = 0$ and hence $\mu[g < f] = 0$ by Theorem 15.2(ii).

Part (iii) for nonnegative f and g follows from part (ii) together with Theorem 10.4. For the general case, prove that $f^+ + g^- = f^- + g^+$ almost everywhere. ∎

Densities

Suppose that δ is a nonnegative measurable function and define a measure ν by (Theorem 16.9)

$$(16.11) \qquad \nu(A) = \int_A \delta \, d\mu, \qquad A \in \mathscr{F};$$

δ is not assumed integrable with respect to μ. Many measures arise in this way. Note that $\mu(A) = 0$ implies that $\nu(A) = 0$. Clearly, ν is finite if and only if δ is integrable μ. Another function δ' gives rise to the same ν if $\delta = \delta'$ almost everywhere. On the other hand, $\nu(A) = \int_A \delta' d\mu$ and (16.11) together imply that $\delta = \delta'$ almost everywhere if μ is σ-finite, as follows from Theorem 16.10(i).

The measure ν defined by (16.11) is said to have *density* δ with respect to μ. A density is by definition nonnegative.

Formal substitution $d\nu = \delta \, d\mu$ gives the formulas (16.12) and (16.13).

Theorem 16.11. *If ν has density δ with respect to μ, then*

(16.12) $$\int f\,d\nu = \int f\delta\,d\mu$$

holds for nonnegative f. Moreover, f (not necessarily nonnegative) is integrable with respect to ν if and only if $f\delta$ is integrable with respect to μ, in which case (16.12) and

(16.13) $$\int_A f\,d\nu = \int_A f\delta\,d\mu$$

both hold. For nonnegative f, (16.13) always holds.

Here $f\delta$ is to be taken as 0 if $f = 0$ or if $\delta = 0$; this is consistent with the conventions (15.2). Note that $\nu[\delta = 0] = 0$.

PROOF. If $f = I_A$, then $\int f\,d\nu = \nu(A)$, so that (16.12) reduces to the definition (16.11). If f is a simple nonnegative function, (16.12) then follows by linearity. If f is nonnegative, then $\int f_n\,d\nu = \int f_n\delta\,d\mu$ for the simple functions f_n of Theorem 13.5, and (16.12) follows by a monotone passage to the limit—that is, by Theorem 16.2. Note that both sides of (16.12) may be infinite.

Even if f is not nonnegative, (16.12) applies to $|f|$, whence it follows that f is integrable with respect to ν if and only if $f\delta$ is integrable with respect to μ. And if f is integrable, (16.12) follows from differencing the same result for f^+ and f^-. Replacing f by fI_A leads from (16.12) to (16.13).   ∎

Example 16.9. If $\nu(A) = \mu(A \cap A_0)$, then (16.11) holds with $\delta = I_{A_0}$, and (16.13) reduces to $\int_A f\,d\nu = \int_{A \cap A_0} f\,d\mu$.   ∎

Theorem 16.11 has two features in common with a number of theorems about integration:

(i) The relation in question, (16.12) in this case, in addition to holding for integrable functions, holds for all nonnegative functions—the point being that if one side of the equation is infinite, then so is the other, and if both are finite, then they have the same value. This is useful in checking for integrability in the first place.

(ii) The result is proved first for indicator functions, then for simple functions, then for nonnegative functions, then for integrable functions. In this connection, see Examples 16.4 and 16.5.

The next result is *Scheffé's theorem.*

Theorem 16.12. *Suppose that* $\nu_n(A) = \int_A \delta_n \, d\mu$ *and* $\nu(A) = \int_A \delta \, d\mu$ *for densities* δ_n *and* δ. *If*

(16.14) $\nu_n(\Omega) = \nu(\Omega) < \infty,$ $n = 1, 2, \ldots,$

and if $\delta_n \to \delta$ *except on a set of* μ-*measure* 0, *then*

(16.15) $\displaystyle\sup_{A \in \mathscr{F}} |\nu(A) - \nu_n(A)| \le \int_\Omega |\delta - \delta_n| \, d\mu \to 0.$

PROOF. The inequality in (16.15) of course follows from (16.5). Let $g_n = \delta - \delta_n$. The positive part g_n^+ of g_n converges to 0 except on a set of μ-measure 0. Moreover, $0 \le g_n^+ \le \delta$ and δ is integrable, and so the dominated convergence theorem applies: $\int g_n^+ \, d\mu \to 0$. But $\int g_n \, d\mu = 0$ by (16.14), and therefore

$$\int_\Omega |g_n| \, d\mu = \int_{[g_n \ge 0]} g_n \, d\mu - \int_{[g_n < 0]} g_n \, d\mu$$

$$= 2\int_{[g_n \ge 0]} g_n \, d\mu = 2\int_\Omega g_n^+ \, d\mu \to 0. \qquad \blacksquare$$

A corollary concerning infinite series follows immediately—take μ as counting measure on $\Omega = \{1, 2, \ldots\}$.

Corollary. *If* $\sum_m x_{nm} = \sum_m x_m < \infty$, *the terms being nonnegative, and if* $\lim_n x_{nm} = x_m$ *for each* m, *then* $\lim_n \sum_m |x_{nm} - x_m| = 0$. *If* y_m *is bounded, then* $\lim_n \sum_m y_m x_{nm} = \sum_m y_m x_m$.

Change of Variable

Let $(\Omega, \mathscr{F})$ and $(\Omega', \mathscr{F}')$ be measurable spaces, and suppose that the mapping $T: \Omega \to \Omega'$ is measurable $\mathscr{F}/\mathscr{F}'$. For a measure μ on $\mathscr{F}$, define a measure μT^{-1} on $\mathscr{F}'$ by

(16.16) $\mu T^{-1}(A') = \mu(T^{-1}A'),$ $A' \in \mathscr{F}',$

as at the end of Section 13.

Suppose f is a real function on Ω' that is measurable $\mathscr{F}'$, so that the composition fT is a real function on Ω that is measurable $\mathscr{F}$ (Theorem 13.1(ii)). The change-of-variable formulas are (16.17) and (16.18). If $A' = \Omega'$, the second reduces to the first.

Theorem 16.13. *If f is nonnegative, then*

$$(16.17) \qquad \int_\Omega f(T\omega)\mu(d\omega) = \int_{\Omega'} f(\omega')\mu T^{-1}(d\omega').$$

A function f (not necessarily nonnegative) is integrable with respect to μT^{-1} if and only if fT is integrable with respect to μ, in which case (16.17) and

$$(16.18) \qquad \int_{T^{-1}A'} f(T\omega)\mu(d\omega) = \int_{A'} f(\omega')\mu T^{-1}(d\omega')$$

hold. For nonnegative f, (16.18) always holds.

PROOF. If $f = I_{A'}$, then $fT = I_{T^{-1}A'}$, and so (16.17) reduces to the definition (16.16). By linearity, (16.17) holds for nonnegative simple functions. If f_n are simple functions for which $0 \le f_n \uparrow f$, then $0 \le f_n T \uparrow fT$, and (16.17) follows by the monotone convergence theorem.

An application of (16.17) to $|f|$ establishes the assertion about integrability, and for integrable f, (16.17) follows by decomposition into positive and negative parts. Finally, if f is replaced by $fI_{A'}$, (16.17) reduces to (16.18). ∎

Example 16.10. Suppose that $(\Omega', \mathcal{F}') = (R^1, \mathcal{R}^1)$ and $T = \varphi$ is an ordinary real function, measurable $\mathcal{F}$. If $f(x) = x$, (16.17) becomes

$$(16.19) \qquad \int_\Omega \varphi(\omega)\mu(d\omega) = \int_{R^1} x\mu\varphi^{-1}(dx).$$

If $\varphi = \sum_i x_i I_{A_i}$ is simple, then $\mu\varphi^{-1}$ has mass $\mu(A_i)$ at x_i, and each side of (16.19) reduces to $\sum_i x_i \mu(A_i)$. ∎

Uniform Integrability

If f is integrable, then $|f|I_{[|f|\ge\alpha]}$ goes to 0 almost everywhere as $\alpha \to \infty$ and is dominated by $|f|$, and hence

$$(16.20) \qquad \lim_{\alpha \to \infty} \int_{[|f|\ge\alpha]} |f|\,d\mu = 0.$$

A sequence $\{f_n\}$ is *uniformly integrable* if (16.20) holds uniformly in n:

$$(16.21) \qquad \lim_{\alpha \to \infty} \sup_n \int_{[|f_n|\ge\alpha]} |f_n|\,d\mu = 0.$$

If (16.21) holds and $\mu(\Omega) < \infty$, and if α is large enough that the supremum in (16.21) is less than 1, then

(16.22)
$$\int |f_n| \, d\mu \leq \alpha \mu(\Omega) + 1,$$

and hence the f_n are integrable. On the other hand, (16.21) always holds if the f_n are uniformly bounded, but the f_n need not in that case be integrable if $\mu(\Omega) = \infty$. For this reason the concept of uniform integrability is interesting only for μ finite

If h is the maximum of $|f|$ and $|g|$, then

$$\int_{|f+g| \geq 2\alpha} |f+g| \, d\mu \leq 2\int_{h \geq \alpha} h \, d\mu \leq 2\int_{|f| \geq \alpha} |f| \, d\mu + 2\int_{|g| \geq \alpha} |g| \, d\mu$$

Therefore, if $\{f_n\}$ and $\{g_n\}$ are uniformly integrable, so is $\{f_n + g_n\}$.

Theorem 16.14. *Suppose that $\mu(\Omega) < \infty$ and $f_n \rightarrow f$ almost everywhere.*

(i) *If the f_n are uniformly integrable, then f is integrable and*

(16.23)
$$\int f_n \, d\mu \rightarrow \int f \, d\mu.$$

(ii) *If f and the f_n are nonnegative and integrable, then (16.23) implies that the f_n are uniformly integrable.*

PROOF. If the f_n are uniformly integrable, it follows by (16.22) and Fatou's lemma that f is integrable. Define

$$f_n^{(\alpha)} = \begin{cases} f_n & \text{if } |f_n| < \alpha, \\ 0 & \text{if } |f_n| \geq \alpha, \end{cases} \qquad f^{(\alpha)} = \begin{cases} f & \text{if } |f| < \alpha, \\ 0 & \text{if } |f| \geq \alpha. \end{cases}$$

If $\mu[|f| = \alpha] = 0$, then $f_n^{(\alpha)} \rightarrow f^{(\alpha)}$ almost everywhere, and by the bounded convergence theorem,

(16.24)
$$\int f_n^{(\alpha)} \, d\mu \rightarrow \int f^{(\alpha)} \, d\mu.$$

Since

(16.25)
$$\int f_n \, d\mu - \int f_n^{(\alpha)} \, d\mu = \int_{|f_n| \geq \alpha} f_n \, d\mu$$

and

(16.26)
$$\int f \, d\mu - \int f^{(\alpha)} \, d\mu = \int_{|f| \geq \alpha} f \, d\mu,$$

it follows from (16.24) that

$$\limsup_n \left| \int f_n \, d\mu - \int f d\mu \right| \leq \sup_n \int_{|f_n| \geq \alpha} |f_n| \, d\mu + \int_{|f| \geq \alpha} |f| \, d\mu.$$

And now (16.23) follows from the uniform integrability and the fact that $\mu[|f| = \alpha] = 0$ for all but countably many α.

Suppose on the other hand that (16.23) holds, where f and the f_n are nonnegative and integrable. If $\mu[f = \alpha] = 0$, then (16.24) holds, and from (16.25) and (16.26) follows

(16.27) $$\int_{f_n \geq \alpha} f_n \, d\mu \to \int_{f \geq \alpha} f d\mu.$$

Since f is integrable, there is, for given ϵ, an α such that the limit in (16 27) is less than ϵ and $\mu[f = \alpha] = 0$. But then the integral on the left is less than ϵ for all n exceeding some n_0. Since the f_n are individually integrable, uniform integrability follows (increase α).   ∎

Corollary. *Suppose that $\mu(\Omega) < \infty$. If f and the f_n are integrable, and if $f_n \to f$ almost everywhere, then these conditions are equivalent*:

(i) f_n *are uniformly integrable*;
(ii) $\int |f - f_n| \, d\mu \to 0$;
(iii) $\int |f_n| \, d\mu \to \int |f| \, d\mu$.

PROOF. If (i) holds, then the differences $|f - f_n|$ are uniformly integrable, and since they converge to 0 almost everywhere, (ii) follows by the theorem. And (ii) implies (iii) because $||f| - |f_n|| \leq |f - f_n|$. Finally, it follows from the theorem that (iii) implies (i).   ∎

Suppose that

(16.28) $$\sup_n \int |f_n|^{1+\epsilon} \, d\mu < \infty$$

for a positive ϵ. If K is the supremum here, then

$$\int_{[|f_n| \geq \alpha]} |f_n| \, d\mu \leq \frac{1}{\alpha^\epsilon} \int_{[|f_n| \geq \alpha]} |f_n|^{1+\epsilon} \, d\mu \leq \frac{K}{\alpha^\epsilon},$$

and so $\{f_n\}$ is uniformly integrable.

Complex Functions

A complex-valued function on Ω has the form $f(\omega) = g(\omega) + ih(\omega)$, where g and h are ordinary finite-valued real functions on Ω. It is, by definition, measurable $\mathcal{F}$ if g and h are. If g and h are integrable, then f is by definition integrable, and its integral is of course taken as

(16.29) $$\int (g + ih) \, d\mu = \int g \, d\mu + i \int h \, d\mu.$$

Since $\max\{|g|,|h|\} \le |f| \le |g| + |h|$, f is integrable if and only if $\int |f|\, d\mu < \infty$, just as in the real case.

The linearity equation (16.3) extends to complex functions and coefficients—the proof requires only that everything be decomposed into real and imaginary parts. Consider the inequality (16.4) for the complex case:

$$(16.30) \qquad \left| \int f d\mu \right| \le \int |f|\, d\mu.$$

If $f = g + ih$ and g and h are simple, the corresponding partitions can be taken to be the same ($g = \sum_k x_k I_{A_k}$ and $h = \sum_k y_k I_{A_k}$), and (16.30) follows by the triangle inequality. For the general integrable f, represent g and h as limits of simple functions dominated by $|f|$, and pass to the limit.

The results on integration to the limit extend as well. Suppose that $f_k = g_k + ih_k$ are complex functions satisfying $\sum_k \int |f_k|\, d\mu < \infty$. Then $\sum_k \int |g_k|\, d\mu < \infty$, and so by the corollary to Theorem 16.7, $\sum_k g_k$ is integrable and integrates to $\sum_k \int g_k\, d\mu$. The same is true of the imaginary parts, and hence $\sum_k f_k$ is integrable and

$$(16 31) \qquad \int \sum_k f_k\, d\mu = \sum_k \int f_k\, d\mu.$$

PROBLEMS

16.1. 13.9↑   Suppose that $\mu(\Omega) < \infty$ and f_n are uniformly bounded.

(a) Assume $f_n \to f$ uniformly and deduce $\int f_n\, d\mu \to \int f d\mu$ from (16.5).

(b) Use part (a) and Egoroff's theorem to give another proof of Theorem 16.5.

16.2. Prove that if $0 \le f_n \to f$ almost everywhere and $\int f_n\, d\mu \le A < \infty$, then f is integrable and $\int f d\mu \le A$. (This is essentially the same as Fatou's lemma and is sometimes called by that name.)

16.3. Suppose that f_n are integrable and $\sup_n \int f_n\, d\mu < \infty$. Show that, if $f_n \uparrow f$, then f is integrable and $\int f_n\, d\mu \to \int f d\mu$. This is *Beppo Levi's theorem*.

16.4. (a) Suppose that functions a_n, b_n, f_n converge almost everywhere to functions a, b, f, respectively. Suppose that the first two sequences may be integrated to the limit—that is, the functions are all integrable and $\int a_n\, d\mu \to \int a d\mu$, $\int b_n\, d\mu \to \int b d\mu$. Suppose, finally, that the first two sequences enclose the third: $a_n \le f_n \le b_n$ almost everywhere. Show that the third may be integrated to the limit.

(b) Deduce Lebesgue's dominated convergence theorem from part (a).

16.5. About Theorem 16.8:

(a) Part (i) is local: there can be a different set A for each t_0. Part (ii) can be recast as a local theorem. Suppose that for $\omega \in A$, where A satisfies (16.9),

$f(\omega, t)$ has derivative $f'(\omega, t_0)$ at t_0; suppose further that

(16.32)
$$\left| \frac{f(\omega, t_0 + h) - f(\omega, t_0)}{h} \right| \le g_1(\omega)$$

for $\omega \in A$ and $0 < |h| < \delta$, where δ is independent of ω and g_1 is integrable. Then $\phi'(t_0) = \int f'(\omega, t_0) \mu(d\omega)$.

The natural way to check (16.32), however, is by the mean-value theorem, and this requires (for $\omega \in A$) a derivative throughout a neighborhood of t_0

(b) If μ is Lebesgue measure on the unit interval Ω, $(a, b) = (0, 1)$, and $f(\omega, t) = I_{(0, t)}(\omega)$, then part (i) applies but part (ii) does not. Why? What about (16.32)?

16.6. Suppose that $f(\omega, \cdot)$ is, for each ω, a function on an open set W in the complex plane and that $f(\cdot, z)$ is for z in W measurable $\mathscr{F}$ and integrable. Suppose that A satisfies (16.9), that $f(\omega, \cdot)$ is analytic in W for ω in A, and that for each z_0 in W there is an integrable $g(\cdot, z_0)$ such that $|f'(\omega, z)| \le g(\omega, z_0)$ for all $\omega \in A$ and all z in some neighborhood of z_0. Show that $\int f(\omega, z) \mu(d\omega)$ is analytic in W.

16.7. (a) Show that if $|f_n| \le g$ and g is integrable, then $\{f_n\}$ is uniformly integrable. Compare the hypotheses of Theorems 16.4 and 16.14

(b) On the unit interval with Lebesgue measure, let $f_n = (n/\log n) I_{(0, n^{-1})}$ for $n \ge 3$. Show that the f_n are uniformly integrable (and $\int f_n d\mu \to 0$) although they are not dominated by any integrable g.

(c) Show for $f_n = n I_{(0, n^{-1}0)} - n I_{(n^{-1}, 2n^{-1})}$ that the f_n can be integrated to the limit (Lebesgue measure) even though the f_n are not uniformly integrable.

16.8. Show that if f is integrable, then for each ϵ there is a δ such that $\mu(A) < \delta$ implies $\int_A |f| d\mu < \epsilon$.

16.9. ↑ Suppose that $\mu(\Omega) < \infty$. Show that $\{f_n\}$ is uniformly integrable if and only if (i) $\int |f_n| d\mu$ is bounded and (ii) for each ϵ there is a δ such that $\mu(A) < \delta$ implies $\int_A |f_n| d\mu < \epsilon$ for all n.

16.10. 2.19 16.9 ↑ Assume $\mu(\Omega) < \infty$.

(a) Show by examples that neither of the conditions (i) and (ii) in the preceding problem implies the other.

(b) Show that (ii) implies (i) for all sequences $\{f_n\}$ if and only if μ is nonatomic.

16.11. Let f be a complex-valued function integrating to $re^{i\theta}$, $r \ge 0$. From $\int(|f(\omega)| - e^{-i\theta} f(\omega)) \mu(d\mu) = \int |f| d\mu - r$, deduce (16.30).

16.12. 11.5 ↑ Consider the vector lattice $\mathscr{L}$ and the functional Λ of Problems 11.4 and 11.5. Let μ be the extension (Theorem 11.3) to $\mathscr{F} = \sigma(\mathscr{F}_0)$ of the set function μ_0 on $\mathscr{F}_0$.

(a) Show by (11.7) that for positive x, y_1, y_2 one has $\nu([f > 1] \times (0, x]) = x \mu_0[f > 1] = x \mu[f > 1]$ and $\nu([y_1 < f \le y_2] \times (0, x]) = x \mu[y_1 < f \le y_2]$.

(b) Show that if $f \in \mathscr{L}$, then f is integrable and

$$\Lambda(f) = \int f d\mu.$$

(Consider first the case $f \geq 0$.) This is the *Daniell–Stone representation theorem*.

SECTION 17. THE INTEGRAL WITH RESPECT TO LEBESGUE MEASURE

The Lebesgue Integral on the Line

A real measurable function on the line is *Lebesgue integrable* if it is integrable with respect to Lebesgue measure λ, and its *Lebesgue integral* $\int f d\lambda$ is denoted by $\int f(x) dx$, or, in the case of integration over an interval, by $\int_a^b f(x) dx$. The theory of the preceding two sections of course applies to the Lebesgue integral. It is instructive to compare it with the Riemann integral.

The Riemann Integral

A real function f on an interval $(a, b]$ is by definition[†] *Riemann integrable*, with integral r, if this condition holds: For each ϵ there exists a δ with the property that

$$(17.1) \qquad \left| r - \sum_i f(x_i) \lambda(I_i) \right| < \epsilon$$

if $\{I_i\}$ is any finite partition of $(a, b]$ into subintervals satisfying $\lambda(I_i) < \delta$ and if $x_i \in I_i$ for each i. The Riemann integral for step functions was used in Section 1.

Suppose that f is Borel measurable, and suppose that f is bounded, so that it is Lebesgue integrable. If f is also Riemann integrable, the r of (17.1) must coincide with the Lebesgue integral $\int_a^b f(x) dx$. To see this, first note that letting x_i vary over I_i leads from (17.1) to

$$(17.2) \qquad \left| r - \sum_i \sup_{x \in I_i} f(x) \cdot \lambda(I_i) \right| \leq \epsilon.$$

Consider the simple function g with value $\sup_{x \in I_i} f(x)$ on I_i. Now $f \leq g$, and the sum in (17.2) is the Lebesgue integral of g. By monotonicity of the

[†]For other definitions, see the first problem at the end of the section and the *Note on terminology* following it.

Lebesgue integral, $\int_a^b f(x)\,dx \le \int_a^b g(x)\,dx \le r + \epsilon$. The reverse inequality follows in the same way, and so $\int_a^b f(x)\,dx = r$. Therefore, *the Riemann integral when it exists coincides with the Lebsegue integral.*

Suppose that f is continuous on $[a, b]$. By uniform continuity, for each ϵ there exists a δ such that $|f(x) - f(y)| < \epsilon/(b - a)$ if $|x - y| < \delta$. If $\lambda(I_i) < \delta$ and $x_i \in I_i$, then $g = \sum_i f(x_i) I_{I_i}$ satisfies $|f - g| < \epsilon/(b - a)$ and hence $|\int_a^b f\,dx - \int_a^b g\,dx| < \epsilon$. But this is (17.1) with r replaced (as it must be) by the Lebesgue integral $\int_a^b f\,dx$: *A continuous function on a closed interval is Riemann integrable.*

Example 17.1. If f is the indicator of the set of rationals in $(0, 1]$, then the Lebesgue integral $\int_0^1(x)\,dx$ is 0 because $f = 0$ almost everywhere. But for an arbitrary partition $\{I_i\}$ of $(0, 1]$ into intervals, $\sum_i f(x_i)\lambda(I_i)$ with $x_i \in I_i$ is 1 if each x_i is taken from the rationals and 0 if each x_i is taken from the irrationals. Thus f is not Riemann integrable.   ∎

Example 17.2. For the f of Example 17.1, there exists a g (namely, $g \equiv 0$) such that $f = g$ almost everywhere and g is Riemann integrable. To show that the Lebesgue theory is not reducible to the Riemann theory by the casting out of sets of measure 0, it is of interest to produce an f (bounded and Borel measurable) for which no such g exists.

In Examples 3.1 and 3.2 there were constructed Borel subsets A of $(0, 1]$ such that $0 < \lambda(A) < 1$ and such that $\lambda(A \cap I) > 0$ for each subinterval I of $(0, 1]$. Take $f = I_A$. Suppose that $f = g$ almost everywhere and that $\{I_i\}$ is a decomposition of $(0, 1]$ into subintervals. Since $\lambda(I_i \cap A \cap [f = g]) = \lambda(I_i \cap A) > 0$, it follows that $g(y_i) = f(y_i) = 1$ for some y_i in $I_i \cap A$, and therefore,

$$(17.3) \qquad\qquad \sum_i g(y_i)\lambda(I_i) = 1 > \lambda(A).$$

If g were Riemann integrable, its Riemann integral would coincide with the Lebesgue integrals $\int g\,dx = \int f\,dx = \lambda(A)$, in contradiction to (17.3).   ∎

It is because of their extreme oscillations that the functions in Examples 17.1 and 17.2 fail to be Riemann integrable. (It can be shown that a bounded function on a bounded interval is Riemann integrable if and only if the set of its discontinuities has Lebesgue measure 0.[†]) This cannot happen in the case of the Lebesgue integral of a measurable function: if f fails to be Lebesgue integrable, it is because its positive part or its negative part is too large, not because one or the other is too irregular.

_____
[†]See Problem 17.1.

Example 17.3. It is an important analytic fact that

(17.4)
$$\lim_{t \to \infty} \int_0^t \frac{\sin x}{x} dx = \frac{\pi}{2}.$$

The existence of the limit is simple to prove, because $\int_{(n-1)\pi}^{n\pi} x^{-1} \sin x \, dx$ alternates in sign and its absolute value decreases to 0; the value of the limit will be identified in the next section (Example 18.4). On the other hand, $x^{-1} \sin x$ is not Lebesgue integrable over $(0, \infty)$, because its positive and negative parts integrate to ∞. Within the conventions of the Lebesgue theory, (17.4) thus cannot be written $\int_0^\infty x^{-1} \sin x \, dx = \pi/2$—although such "improper" integrals appear in calculus texts. It is, of course, just a question of choosing the terminology most convenient for the subject at hand.   ∎

The function in Example 17.2 is not equal almost everywhere to any Riemann integrable function. Every Lebesgue integrable function can, however, be approximated in a certain sense by Riemann integrable functions of two kinds.

Theorem 17.1. *Suppose that $\int |f| \, dx < \infty$ and $\epsilon > 0$.*

(i) *There is a step function $g = \sum_{i=1}^k x_i I_{A_i}$, with bounded intervals as the A_i, such that $\int |f - g| \, dx < \epsilon$.*

(ii) *There is a continuous integrable h with bounded support such that $\int |f - h| \, dx < \epsilon$.*

PROOF. By the construction (13.6) and the dominated convergence theorem, (i) holds if the A_i are not required to be intervals; moreover, $\lambda(A_i) < \infty$ for each i for which $x_i \neq 0$. By Theorem 11.4 there exists a finite disjoint union B_i of intervals such that $\lambda(A_i \Delta B_i) < \epsilon/k|x_i|$. But then $\sum_i x_i I_{B_i}$ satisfies the requirements of (i) with 2ϵ in place of ϵ.

To prove (ii) it is only necessary to show that for the g of (i) there is a continuous h such that $\int |g - h| \, dx < \epsilon$. Suppose that $A_i = (a_i, b_i]$; let $h_i(x)$ be 1 on $(a_i, b_i]$ and 0 outside $(a_i - \delta, b_i + \delta]$, and let it increase linearly from 0 to 1 over $(a_i - \delta, a_i]$ and decrease linearly from 1 to 0 over $(b_i, b_i + \delta]$. Since $\int |I_{A_i} - h_i| \, dx \to 0$ as $\delta \to 0$, $h = \sum x_i h_i$ for sufficiently small δ will satisfy the requirements.   ∎

The Lebesgue integral is thus determined by its values for continuous functions.[†]

[†]This provides another way of *defining* the Lebesgue integral on the line. See Problem 17.13.

The Fundamental Theorem of Calculus

Adopt the convention that $\int_\alpha^\beta = -\int_\beta^\alpha$ if $\alpha > \beta$. For positive h,

$$\left| \frac{1}{h} \int_x^{x+h} f(y)\, dy - f(x) \right| \le \frac{1}{h} \int_x^{x+h} |f(y) - f(x)|\, dy$$

$$\le \sup \big[|f(y) - f(x)| \colon x \le y \le x + h \big],$$

and the right side goes to 0 with h if f is continuous at x. The same thing holds for negative h, and therefore $\int_a^x f(y)\, dy$ has derivative $f(x)$:

$$(17.5) \qquad\qquad \frac{d}{dx} \int_a^x f(y)\, dy = f(x)$$

if f is continuous at x.

Suppose that F is a function with continuous derivative $F' = f$; that is, suppose that F is a *primitive* of the continuous function f. Then

$$(17.6) \qquad\qquad \int_a^b f(x)\, dx = \int_a^b F'(x)\, dx = F(b) - F(a),$$

as follows from the fact that $F(x) - F(a)$ and $\int_a^x f(y)\, dy$ agree at $x = a$ and by (17.5) have identical derivatives. For continuous f, (17.5) and (17.6) are two ways of stating the fundamental theorem of calculus. To the calculation of Lebesgue integrals the methods of elementary calculus thus apply.

As will follow from the general theory of derivatives in Section 31, (17.5) holds outside a set of Lebesgue measure 0 if f is integrable—it need not be continuous. As the following example shows, however, (17.6) can fail for discontinuous f.

Example 17.4. Define $F(x) = x^2 \sin x^{-2}$ for $0 \le x \le \frac{1}{2}$ and $F(x) = 0$ for $x \le 0$ and for $x \ge 1$. Now for $\frac{1}{2} < x < 1$ define $F(x)$ in such a way that F is continuously differentiable over $(0, \infty)$. Then F is everywhere differentiable, but $F'(0) = 0$ and $F'(x) = 2x \sin x^{-2} - 2x^{-1} \cos x^{-2}$ for $0 < x < \frac{1}{2}$. Thus F' is discontinuous at 0; F' is, in fact, not even integrable over $(0, 1]$, which makes (17.6) impossible for $a = 0$.

For a more extreme example, decompose $(0, 1]$ into countably many subintervals $(a_n, b_n]$. Define $G(x) = 0$ for $x \le 0$ and $x \ge 1$, and on $(a_n, b_n]$ define $G(x) = F((x - a_n)/(b_n - a_n))$. Then G is everywhere differentiable, but (17.6) is impossible for G if $(a, b]$ contains any of the $(a_n, b_n]$, because G is not integrable over any of them. ∎

Change of Variable

For

$$(17.7) \qquad\qquad [a, b] \xrightarrow{T} [u, v] \xrightarrow{f} R^1,$$

the change-of-variable formula is

$$(17.8) \qquad \int_a^b f(Tx)T'(x)\, dx = \int_{Ta}^{Tb} f(y)\, dy.$$

If T' exists and is continuous, and if f is continuous, the two integrals are finite because the integrands are bounded, and to prove (17.8) it is enough to let b be a variable and differentiate with respect to it.[†]

With the obvious limiting arguments, this applies to unbounded intervals and to open ones:

Example 17.5. Put $T(x) = \tan x$ on $(-\pi/2, \pi/2)$. Then $T'(x) = 1 + T^2(x)$, and (17.8) for $f(y) = (1 + y^2)^{-1}$ gives

$$(17.9) \qquad \int_{-\infty}^{\infty} \frac{dy}{1 + y^2} = \pi. \qquad\qquad\qquad\blacksquare$$

The Lebesgue Integral in R^k

The k-dimensional Lebesgue integral, the integral in $(R^k, \mathscr{R}^k, \lambda_k)$, is denoted $\int f(x)\, dx$, it being understood that $x = (x_1, \ldots, x_k)$. In low-dimensional cases it is also denoted $\int\int_A f(x_1, x_2)\, dx_1\, dx_2$, and so on.

As for the rule for changing variables, suppose that $T: U \to R^k$, where U is an open set in R^k. The map has the form $Tx = (t_1(x), \ldots, t_k(x))$; it is by definition continuously differentiable if the partial derivatives $t_{ij}(x) = \partial t_i/\partial x_j$ exist and are continuous in U. Let $D_x = [t_{ij}(x)]$ be the Jacobian matrix, let $J(x) = \det D_x$ be the Jacobian determinant, and let $V = TU$.

Theorem 17.2. *Let T be a continuously differentiable map of the open set U onto V. Suppose that T is one-to-one and that $J(x) \neq 0$ for all x. If f is nonnegative, then*

$$(17.10) \qquad \int_U f(Tx)|J(x)|\, dx = \int_V f(y)\, dy.$$

By the inverse–function theorem [A35], V is open and the inverse point mapping T^{-1} is continuously differentiable. It is assumed in (17.10) that f: $V \to R^1$ is a Borel function. As usual, for the general f, (17.10) holds with $|f|$ in place of f, and if the two sides are finite, the absolute-value bars can be removed; and of course f can be replaced by fI_B or fI_{TA}.

[†]See Problem 17.11 for extensions.

Example 17.6. Suppose that T is a nonsingular linear transformation on $U = V = R^k$. Then D_x is for each x the matrix of the transformation. If T is identified with this matrix, then (17.10) becomes

$$(17.11) \qquad |\det T| \int_U f(Tx) \, dx = \int_V f(y) \, dy.$$

If $f = I_{TA}$, this holds because of (12.2), and then it follows in the usual sequence for simple f and for the general nonnegative f: Theorem 17.2 is easy in the linear case. ∎

Example 17.7. In R^2 take $U = \{(\rho, \theta): \rho > 0, \, 0 < \theta < 2\pi\}$ and $T(\rho, \theta) = (\rho \cos \theta, \rho \sin \theta)$. The Jacobian is $J(\rho, \theta) = \rho$, and (17.10) gives the formula for integrating in polar coordinates:

$$(17.12) \qquad \iint_{\substack{\rho > 0 \\ 0 < \theta < 2\pi}} f(\rho \cos \theta, \rho \sin \theta) \rho \, d\rho \, d\theta = \iint_{R^2} f(x, y) \, dx dy.$$

Here V is R^2 with the ray $\{(x, 0): x \geq 0\}$ removed; (17.12) obviously holds even though the ray is included on the right. If the constraint on θ is replaced by $0 < \theta < 4\pi$, for example, then (17.12) is false (a factor of 2 is needed on the right). This explains the assumption that T is one-to-one. ∎

Theorem 17.2 is not the strongest possible; it is only necessary to assume that T is one-to-one on the set $U_0 = [x \in U: J(x) \neq 0]$. This is because, by Sard's theorem,[†] $\lambda_k(V - TU_0) = 0$.

PROOF OF THEOREM 17.2. Suppose it is shown that

$$(17.13) \qquad \int_U f(Tx)|J(x)| \, dx \geq \int_V f(y) \, dy$$

for nonnegative f. Apply this to $T^{-1}: V \to U$, which [A35] is continuously differentiable and has Jacobian $J^{-1}(T^{-1}y)$ at y:

$$\int_V g(T^{-1}y)|J^{-1}(T^{-1}y)| \, dy \geq \int_U g(x) \, dx$$

for nonnegative g on V. Taking $g(x) = f(Tx)|J(x)|$ here leads back to (17.13), but with the inequality reversed. Therefore, proving (17.13) will be enough.

For $f = I_{TA}$, (17.13) reduces to

$$(17.14) \qquad \int_A |J(x)| \, dx \geq \lambda_k(TA).$$

[†]SPIVAK, p. 72.

Each side of (17.14) is a measure on $\mathscr{U} = U \cap \mathscr{R}^k$. If $\mathscr{A}$ consists of the rectangles A satisfying $A^- \subset U$, then $\mathscr{A}$ is a semiring generating $\mathscr{U}$, U is a countable union of $\mathscr{A}$-sets, and the left side of (17.14) is finite for A in $\mathscr{A}$ ($\sup_A -|J| < \infty$). It follows by Corollary 2 to Theorem 11.4 that if (17.14) holds for A in $\mathscr{A}$, then it holds for A in $\mathscr{U}$. But then (linearity and monotone convergence) (17.13) will follow.

Proof of (17.14) *for A in* $\mathscr{A}$. Split the given rectangle A into finitely many subrectangles Q_i satisfying

$$(17.15) \qquad\qquad\qquad \operatorname{diam} Q_i < \delta,$$

δ to be determined. Let x_i be some point of Q_i. Given ϵ, choose δ in the first place so that $|J(x) - J(x')| < \epsilon$ if $x, x' \in A^-$ and $|x - x'| < \delta$. Then (17.15) implies

$$(17.16) \qquad\qquad \sum_i |J(x_i)| \lambda_k(Q_i) \le \int_A |J(x)| \, dx + \epsilon \lambda_k(A).$$

Let $Q_i^{+\epsilon}$ be a rectangle that is concentric with Q_i and similar to it and whose edge lengths are those of Q_i multiplied by $1 + \epsilon$. For x in U consider the affine transformation

$$(17.17) \qquad\qquad \phi_x z = D_x(z - x) + Tx, \qquad z \in R^k;$$

$\phi_x z$ will [A34] be a good approximation to Tz for z near x. Suppose, as will be proved in a moment, that for each ϵ there is a δ such that, if (17.15) holds, then, for each i, ϕ_{x_i} approximates T so well on Q_i that

$$(17.18) \qquad\qquad\qquad TQ_i \subset \phi_{x_i} Q_i^{+\epsilon}.$$

By Theorem 12.2, which shows in the nonsingular case how an affine transformation changes the Lebesgue measures of sets, $\lambda_k(\phi_{x_i} Q_i^{+\epsilon}) = |J(x_i)| \lambda_k(Q_i^{+\epsilon})$. If (17.18) holds, then

$$(17.19) \qquad \lambda_k(TA) = \sum_i \lambda_k(TQ_i) \le \sum_i \lambda_k(\phi_{x_i} Q_i^{+\epsilon})$$

$$= \sum_i |J(x_i)| \lambda_k(Q_i^{+\epsilon}) = (1 + \epsilon)^k \sum_i |J(x_i)| \lambda_k(Q_i).$$

(This, the central step in the proof, shows where the Jacobian in (17.10) comes from.) If for each ϵ there is a δ such that (17.15) implies both (17.16) and (17.19), then (17.14) will follow. Thus everything depends on (17.18), and the remaining problem is to show that for each ϵ there is a δ such that (17.18) holds if (17.15) does.

Proof of (17.18). As (x, z) varies over the compact set $A^- \times [z: |z| = 1]$, $|D_x^{-1} z|$ is continuous, and therefore, for some c,

$$(17.20) \qquad\qquad |D_x^{-1} z| \le c|z| \qquad \text{for } x \in A, z \in R^k.$$

Since the t_{jl} are uniformly continuous on A^-, δ can be chosen so that $|t_{jl}(z) - t_{jl}(x)| \le \epsilon/k^2 c$ for all j, l if $z, x \in A$ and $|z - x| < \delta$. But then, by linear approximation [A34: (16)], $|Tz - Tx - D_x(z - x)| \le \epsilon c^{-1}|z - x| < \epsilon c^{-1} \delta$. If (17.15) holds and $\delta < 1$, then by the definition (17.17),

$$(17.21) \qquad\qquad |Tz - \phi_{x_i} z| < \epsilon/c \qquad \text{for } z \in Q_i.$$

To prove (17.18), note that $z \in Q_i$ implies

$$|\phi_{x_i}^{-1}Tz - z| = |\phi_{x_i}^{-1}Tz - \phi_{x_i}^{-1}\phi_{x_i}z| = |D_{x_i}^{-1}(Tz - \phi_{x_i}z)|$$

$$\leq c|Tz - \phi_{x_i}z| < \epsilon,$$

where the first inequality follows by (17.20) and the second by (17.21). Since $\phi_{x_i}^{-1}Tz$ is within ϵ of the point z of Q_i, it lies in $Q_i^{+\epsilon}$: $\phi_{x_i}^{-1}Tz \in Q_i^{+\epsilon}$, or $Tz \in \phi_{x_i}Q_i^{+\epsilon}$. Hence (17.18) holds, which completes the proof.   ∎

Stieltjes Integrals

Suppose that F is a function on R^k satisfying the hypotheses of Theorem 12.5, so that there exists a measure μ such that $\mu(A) = \Delta_A F$ for bounded rectangles A. In integrals with respect to μ, $\mu(dx)$ is often replaced by $dF(x)$:

$$(17.22) \qquad \int_A f(x)\, dF(x) = \int_A f(x)\mu(dx).$$

The left side of this equation is the *Stieltjes integral* of f with respect to F; since it is defined by the right side of the equation, nothing new is involved.

Suppose that f is uniformly continuous on a rectangle A, and suppose that A is decomposed into rectangles A_m small enough that $|f(x) - f(y)| < \epsilon/\mu(A)$ for $x, y \in A_m$. Then

$$\left| \int_A f(x)\, dF(x) - \sum_m f(x_m)\Delta_{A_m}F \right| < \epsilon$$

for $x_m \in A_m$. In this case the left side of (17.22) can be defined as the limit of these approximating sums without any reference to the general theory of measure, and for historical reasons it is sometimes called the *Riemann–Stieltjes* integral; (17.22) for the general f is then called the *Lebesgue–Stieltjes* integral. Since these distinctions are unimportant in the context of general measure theory, $\int f(x)\, dF(x)$ and $\int f dF$ are best regarded as merely notational variants for $\int f(x)\mu(dx)$ and $\int f d\mu$.

PROBLEMS

Let f be a bounded function on a bounded interval, say $[0, 1]$. Do not assume that f is a Borel function. Denote by $L_* f$ and $L^* f$ (L for Lebesgue) the lower and upper integrals as defined by (15.9) and (15.10), where μ is now Lebesgue measure λ on the Borel sets of $[0, 1]$. Denote by $R_* f$ and $R^* f$ (R for Riemann) the same quantities but with the outer supremum and infimum in (15.9) and (15.10) extending only over finite partitions of $[0, 1]$ into subintervals. It is obvious (see (15.11)) that

$$(17.23) \qquad R_* f \leq L_* f \leq L^* f \leq R^* f.$$

Suppose that f is bounded, and consider these three conditions:

(i) *There is an r with the property that for each ϵ there is a δ such that* (17.1) *holds if $\{I_i\}$ partitions $[0,1]$ into subintervals with $\lambda(I_i) < \delta$ and if $x_i \in I_i$.*

(ii) $R_* f = R^* f$.

(iii) *If D_f is the set of points of discontinuity of f, then $\lambda(D_f) = 0$.*

The conditions are equivalent.

17.1. Prove:

(a) D_f is a Borel set.

(b) (i) implies (ii).

(c) (ii) implies (iii).

(d) (iii) implies (i).

(e) The r of (i) must coincide with the $R_* f = R^* f$ of (ii).

A note on terminology. An f on the general $(\Omega, \mathscr{F}, \mu)$ is defined to be integrable not if (15.12) holds, but if (16.1) does. And an f on $[0,1]$ is defined to be integrable with respect to Lebesgue measure not if $L_* f = L^* f$ holds, but, rather, if

$$(17.24) \qquad \int_0^1 |f(x)| \, dx < \infty$$

does. The condition $L_* f = L^* f$ is not at issue, since for bounded f it always holds if f is a Borel function, and in this book f is always assumed to be a Borel function unless the contrary is explicitly stated. For the Lebesgue integral, the question is whether f is small enough that (17.24) holds, not whether it is sufficiently regular that $L_* f = L^* f$. For the Riemann integral, the terminology is different because $R_* f < R^* f$ holds for all sorts of important Borel functions, and one way to define Riemann integrability is to require $R_* f = R^* f$. In the context of general integration theory, one occasionally looks at the Riemann integral, but mostly for illustration and comparison.

17.2. 3.15 17.1↑ (a) Show that an indicator I_A for $A \subset [0,1]$ is Riemann integrable if and only if A is Jordan measurable.

(b) Find a Riemann integrable function that is not a Borel function.

17.3. Extend Theorem 17.1 to R^k.

17.4. Show that if f is integrable, then

$$\lim_{t \to 0} \int |f(x+t) - f(x)| \, dx = 0.$$

Use Theorem 17.1.

17.5. Suppose that μ is a finite measure on $\mathscr{R}^k$ and A is closed. Show that $\mu(x + A)$ is upper semicontinuous in x and hence measurable.

17.6. Suppose that $\int_0^\infty |f(x)| \, dx < \infty$. Show that for each ϵ, $\lambda[x : x > \alpha, |f(x)| > \epsilon] \to 0$ as $\alpha \to \infty$. Show by example that $f(x)$ need not go to 0 as $x \to \infty$ (even if f is continuous).

17.7. Let $f_n(x) = x^{n-1} - 2x^{2n-1}$. Calculate and compare $\int_0^1 \sum_{n=1}^\infty f_n(x)\,dx$ and $\sum_{n=1}^\infty \int_0^1 f_n(x)\,dx$. Relate this to Theorem 16.6 and to the corollary to Theorem 16.7.

17.8. Show that $(1+y^2)^{-1}$ has equal integrals over $(-\infty,-1),(-1,0),(0,1),(1,\infty)$. Conclude from (17.9) that $\int_0^1 (1+y^2)^{-1}\,dy = \pi/4$. Expand the integrand in a geometric series and deduce Leibniz's formula

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots$$

by Theorem 16.7 (note that its corollary does not apply).

17.9. Show that if f is integrable, there exist continuous, integrable functions g_n such that $g_n(x) \to f(x)$ except on a set of Lebesgue measure 0. (Use Theorem 17.1(ii) with $\epsilon = n^{-2}$.)

17.10. 13.9 17.9↑ Let f be a finite-valued Borel function over $[0,1]$ By the following steps, prove *Lusin's theorem*: For each ϵ there exists a continuous function g such that $\lambda[x \in (0,1): f(x) \neq g(x)] < \epsilon$.

(a) Show that f may be assumed integrable, or even bounded.

(b) Let g_n be continuous functions converging to f almost everywhere. Combine Egoroff's theorem and Theorem 12.3 to show that convergence is uniform on a compact set K such that $\lambda((0,1) - K) < \epsilon$. The limit $\lim_n g_n(x) = f(x)$ must be continuous when restricted to K.

(c) Exhibit $(0,1) - K$ as a disjoint union of open intervals I_k [A12], define g as f on K, and define it by linear interpolation on each I_k.

17.11. Suppose in (17.7) that T' exists and is continuous and f is a Borel function, and suppose that $\int_a^b |f(Tx)T'(x)|\,dx < \infty$. Show in steps that $\int_{T(a,b)} |f(y)|\,dy < \infty$ and (17.8) holds. Prove this for (a) f continuous, (b) $f = I_{[s,t]}$, (c) $f = I_B$, (d) f simple, (e) $f \geq 0$, (f) f general.

17.12. 16.12↑ Let $\mathscr{L}$ consist of the continuous functions on R^1 with compact support. Show that $\mathscr{L}$ is a vector lattice in the sense of Problem 11.4 and has the property that $f \in \mathscr{L}$ implies $f \wedge 1 \in \mathscr{L}$ (note that $1 \notin \mathscr{L}$). Show that the σ-field $\mathscr{F}$ generated by $\mathscr{L}$ is $\mathscr{R}^1$. Suppose Λ is a positive linear functional on $\mathscr{L}$; show that Λ has the required continuity property if and only if $f_n(x) \downarrow 0$ uniformly in x implies $\Lambda(f_n) \to 0$. Show under this assumption on Λ that there is a measure μ on $\mathscr{R}^1$ such that

$$(17.25) \qquad \Lambda(f) = \int f\,d\mu, \qquad f \in \mathscr{L}.$$

Show that μ is σ-finite and unique. This is a version of the *Riesz representation theorem*.

17.13. ↑ Let $\Lambda(f)$ be the Riemann integral of f, which does exist for f in $\mathscr{L}$. Using the most elementary facts about Riemann integration, show that the μ determined by (17.25) is Lebesgue measure. This gives still another way of constructing Lebesgue measure.

17.14. ↑ Extend the ideas in the preceding two problems to R^k.

SECTION 18. PRODUCT MEASURE AND FUBINI'S THEOREM

Let $(X, \mathscr{X})$ and $(Y, \mathscr{Y})$ be measurable spaces. For given measures μ and ν on these spaces, the problem is to construct on the Cartesian product $X \times Y$ a *product measure* π such that $\pi(A \times B) = \mu(A)\nu(B)$ for $A \subset X$ and $B \subset Y$. In the case where μ and ν are Lebesgue measure on the line, π will be Lebesgue measure in the plane. The main result is *Fubini's theorem*, according to which double integrals can be calculated as iterated integrals.

Product Spaces

It is notationally convenient in this section to change from $(\Omega, \mathscr{F})$ to $(X, \mathscr{X})$ and $(Y, \mathscr{Y})$. In the product space $X \times Y$ a *measurable rectangle* is a product $A \times B$ for which $A \in \mathscr{X}$ and $B \in \mathscr{Y}$. The natural class of sets in $X \times Y$ to consider is the σ-field $\mathscr{X} \times \mathscr{Y}$ generated by the measurable rectangles. (Of course, $\mathscr{X} \times \mathscr{Y}$ is not a Cartesian product in the usual sense.)

Example 18.1. Suppose that $X = Y = R^1$ and $\mathscr{X} = \mathscr{Y} = \mathscr{R}^1$. Then a measurable rectangle is a Cartesian product $A \times B$ in which A and B are linear Borel sets. The term *rectangle* has up to this point been reserved for Cartesian products of intervals, and so a measurable rectangle is more general. As the measurable rectangles do include the ordinary ones and the latter generate $\mathscr{R}^2$, it follows that $\mathscr{R}^2 \subset \mathscr{R}^1 \times \mathscr{R}^1$. On the other hand, if A is an interval, $[B: A \times B \in \mathscr{R}^2]$ contains R^1 $(A \times R^1 = \bigcup_n (A \times (-n, n]) \in \mathscr{R}^2)$ and is closed under the formation of proper differences and countable unions; thus it is a σ-field containing the intervals and hence the Borel sets. Therefore, if B is a Borel set, $[A: A \times B \in \mathscr{R}^2]$ contains the intervals and hence, being a σ-field, contains the Borel sets. Thus all the measurable rectangles are in $\mathscr{R}^2$, and so $\mathscr{R}^1 \times \mathscr{R}^1 = \mathscr{R}^2$ consists exactly of the two-dimensional Borel sets. ∎

As this example shows, $\mathscr{X} \times \mathscr{Y}$ is in general much larger than the class of measurable rectangles.

Theorem 18.1. (i) *If $E \in \mathscr{X} \times \mathscr{Y}$, then for each x the set $[y: (x, y) \in E]$ lies in $\mathscr{Y}$ and for each y the set $[x: (x, y) \in E]$ lies in $\mathscr{X}$.*

(ii) *If f is measurable $\mathscr{X} \times \mathscr{Y}$, then for each fixed x the function $f(x, \cdot)$ is measurable $\mathscr{Y}$, and for each fixed y the function $f(\cdot, y)$ is measurable $\mathscr{X}$.*

The set $[y: (x, y) \in E]$ is the *section* of E determined by x, and $f(x, \cdot)$ is the *section* of f determined by x.

PROOF. Fix x, and consider the mapping $T_x: Y \to X \times Y$ defined by $T_x y = (x, y)$. If $E = A \times B$ is a measurable rectangle, $T_x^{-1} E$ is B or $\varnothing$

according as A contains x or not, and in either case $T_x^{-1}E \in \mathcal{Y}$. By Theorem 13.1(i), T_x is measurable $\mathcal{Y}/\mathcal{X} \times \mathcal{Y}$. Hence $[y: (x,y) \in E] = T_x^{-1}E \in \mathcal{Y}$ for $E \in \mathcal{X} \times \mathcal{Y}$. By Theorem 13.1(ii), if f is measurable $\mathcal{X} \times \mathcal{Y}/\mathcal{R}^1$, then fT_x is measurable $\mathcal{Y}/\mathcal{R}^1$. Hence $f(x,\cdot) = fT_x(\cdot)$ is measurable $\mathcal{Y}$. The symmetric statements for fixed y are proved the same way.   ∎

Product Measure

Now suppose that $(X, \mathcal{X}, \mu)$ and $(Y, \mathcal{Y}, \nu)$ are measure spaces, and suppose for the moment that μ and ν are *finite*. By the theorem just proved $\nu[y: (x, y) \in E]$ is a well-defined function of x. If $\mathcal{L}$ is the class of E in $\mathcal{X} \times \mathcal{Y}$ for which this function is measurable $\mathcal{X}$, it is not hard to show that $\mathcal{L}$ is a λ-system. Since the function is $I_A(x)\nu(B)$ for $E = A \times B$, $\mathcal{L}$ contains the π-system consisting of the measurable rectangles. Hence $\mathcal{L}$ coincides with $\mathcal{X} \times \mathcal{Y}$ by the π-λ theorem. It follows without difficulty that

$$(18.1) \qquad \pi'(E) = \int_X \nu[y: (x, y) \in E]\mu(dx), \qquad E \in \mathcal{X} \times \mathcal{Y},$$

is a finite measure on $\mathcal{X} \times \mathcal{Y}$, and similarly for

$$(18.2) \qquad \pi''(E) = \int_Y \mu[x: (x, y) \in E]\nu(dy), \qquad E \in \mathcal{X} \times \mathcal{Y}.$$

For measurable rectangles,

$$(18.3) \qquad \pi'(A \times B) = \pi''(A \times B) = \mu(A) \cdot \nu(B).$$

The class of E in $\mathcal{X} \times \mathcal{Y}$ for which $\pi'(E) = \pi''(E)$ thus contains the measurable rectangles; since this class is a λ-system, it contains $\mathcal{X} \times \mathcal{Y}$. The common value $\pi'(E) = \pi''(E)$ is the product measure sought.

To show that (18.1) and (18.2) also agree for σ-finite μ and ν, let $\{A_m\}$ and $\{B_n\}$ be decompositions of X and Y into sets of finite measure, and put $\mu_m(A) = \mu(A \cap A_m)$ and $\nu_n(B) = \nu(B \cap B_n)$. Since $\nu(B) = \Sigma_m \nu_m(B)$, the integrand in (18.1) is measurable $\mathcal{X}$ in the σ-finite as well as in the finite case; hence π' is a well-defined measure on $\mathcal{X} \times \mathcal{Y}$ and so is π''. If π'_{mn} and π''_{mn} are (18.1) and (18.2) for μ_m and ν_n, then by the finite case, already treated (see Example 16.5), $\pi'(E) = \Sigma_{mn}\pi'_{mn}(E) = \Sigma_{mn}\pi''_{mn}(E) = \pi''(E)$. Thus (18.1) and (18.2) coincide in the σ-finite case as well. Moreover, $\pi'(A \times B) = \Sigma_{mn}\mu_m(A)\nu_n(B) = \mu(A)\nu(B)$.

Theorem 18.2. *If $(X, \mathcal{X}, \mu)$ and $(Y, \mathcal{Y}, \nu)$ are σ-finite measure spaces, $\pi(E) = \pi'(E) = \pi''(E)$ defines a σ-finite measure on $\mathcal{X} \times \mathcal{Y}$; it is the only measure such that $\pi(A \times B) = \mu(A) \cdot \nu(B)$ for measurable rectangles.*

PROOF. Only σ-finiteness and uniqueness remain to be proved. The products $A_m \times B_n$ for $\{A_m\}$ and $\{B_n\}$ as above decompose $X \times Y$ into measurable rectangles of finite π-measure. This proves both σ-finiteness and uniqueness, since the measurable rectangles form a π-system generating $\mathscr{X} \times \mathscr{Y}$ (Theorem 10.3).   ■

The π thus defined is called *product measure*; it is usually denoted $\mu \times \nu$. Note that the integrands in (18.1) and (18.2) may be infinite for certain x and y, which is one reason for introducing functions with infinite values. Note also that (18.3) in some cases requires the conventions (15.2).

Fubini's Theorem

Integrals with respect to π are usually computed via the formulas

$$(18.4) \qquad \int_{X \times Y} f(x, y) \pi(d(x, y)) = \int_X \left[\int_Y f(x, y) \nu(dy) \right] \mu(dx)$$

and

$$(18.5) \qquad \int_{X \times Y} f(x, y) \pi(d(x, y)) = \int_Y \left[\int_X f(x, y) \mu(dx) \right] \nu(dy).$$

The left side here is a *double* integral, and the right sides are *iterated* integrals. The formulas hold very generally, as the following argument shows.

Consider (18.4). The inner integral on the right is

$$(18.6) \qquad \int_Y f(x, y) \nu(dy).$$

Because of Theorem 18.1(ii), for f measurable $\mathscr{X} \times \mathscr{Y}$ the integrand here is measurable $\mathscr{Y}$; the question is whether the integral exists, whether (18.6) is measurable $\mathscr{X}$ as a function of x, and whether it integrates to the left side of (18.4).

First consider nonnegative f. If $f = I_E$, everything follows from Theorem 18.2: (18.6) is $\nu[y: (x, y) \in E]$, and (18.4) reduces to $\pi(E) = \pi'(E)$. Because of linearity (Theorem 15.1(iv)), if f is a nonnegative simple function, then (18.6) is a linear combination of functions measurable $\mathscr{X}$ and hence is itself measurable $\mathscr{X}$; further application of linearity to the two sides of (18.4) shows that (18.4) again holds. The general nonnegative f is the monotone limit of nonnegative simple functions; applying the monotone convergence theorem to (18.6) and then to each side of (18.4) shows that again f has the properties required.

Thus for nonnegative f, (18.6) is a well-defined function of x (the value ∞ is not excluded), measurable $\mathscr{X}$, whose integral satisfies (18.4). If one side of

(18.4) is infinite, so is the other; if both are finite, they have the same finite value.

Now suppose that f, not necessarily nonnegative, is integrable with respect to π. Then the two sides of (18.4) are finite if f is replaced by $|f|$. Now make the further assumption that

$$(18.7) \qquad \int_Y |f(x, y)| \nu(dy) < \infty$$

for all x. Then

$$(18.8) \qquad \int_Y f(x, y)\nu(dy) = \int_Y f^+(x, y)\nu(dy) - \int_Y f^-(x, y)\nu(dy).$$

The functions on the right here are measurable $\mathscr{X}$ and (since $f^+, f^- \le |f|$) integrable with respect to μ, and so the same is true of the function on the left. Integrating out the x and applying (18.4) to f^+ and to f^- gives (18.4) for f itself.

The set A_0 of x satisfying (18.7) need not coincide with X, but $\mu(X - A_0) = 0$ if f is integrable with respect to π, because the function in (18.7) integrates to $\int |f| d\pi$ (Theorem 15.2(iii)). Now (18.8) holds on A_0, (18.6) is measurable $\mathscr{X}$ on A_0, and (18.4) again follows if the inner integral on the right is given some arbitrary constant value on $X - A_0$.

The same analysis applies to (18.5):

Theorem 18.3. *Under the hypotheses of Theorem* 18.2, *for nonnegative* f *the functions*

$$(18.9) \qquad \int_Y f(x, y)\nu(dy), \int_X f(x, y)\mu(dx)$$

are measurable $\mathscr{X}$ *and* $\mathscr{Y}$, *respectively, and* (18.4) *and* (18.5) *hold. If* f (*not necessarily nonnegative*) *is integrable with respect to* π, *then the two functions* (18.9) *are finite and measurable on* A_0 *and on* B_0, *respectively, where* $\mu(X - A_0) = \nu(Y - B_0) = 0$, *and again* (18.4) *and* (18.5) *hold.*

It is understood here that the inner integrals on the right in (18.4) and (18.5) are set equal to 0 (say) outside A_0 and B_0.[†]

This is *Fubini's theorem*; the part concerning nonnegative f is sometimes called *Tonelli's theorem*. Application of the theorem usually follows a two-step procedure that parallels its proof. First, one of the iterated integrals is computed (or estimated above) with $|f|$ in place of f. If the result is finite,

---

[†]Since two functions that are equal almost everywhere have the same integral, the theory of integration could be extended to functions that are only *defined* almost everywhere; then A_0 and B_0 would disappear from Theorem 18.3.

then the double integral (integral with respect to π) of $|f|$ must be finite, so that f is integrable with respect to π; then the value of the double integral of f is found by computing one of the iterated integrals of f. If the iterated integral of $|f|$ is infinite, f is not integrable π.

Example 18.2. Let D_r be the closed disk in the plane with center at the origin and radius r. By (17.12),

$$\lambda_2(D_r) = \iint_{D_r} dx\,dy = \iint_{\substack{0 < p \leq r \\ 0 < 0 < 2\pi}} \rho\,d\rho\,d\theta.$$

The last integral can be evaluated by Fubini's theorem:

$$\lambda_2(D_r) = 2\pi \int_0^r \rho\,d\rho = \pi r^2. \qquad \blacksquare$$

Example 18.3. Let $I = \int_{-\infty}^{\infty} e^{-x^2} dx$. By Fubini's theorem applied in the plane and by the polar-coordinate formula,

$$I^2 = \iint_{R^2} e^{-(x^2+y^2)} dx\,dy = \iint_{\substack{p > 0 \\ 0 < \theta < 2\pi}} e^{-\rho^2} \rho\,d\rho\,d\theta.$$

The double integral on the right can be evaluated as an iterated integral by another application of Fubini's theorem, which leads to the famous formula

(18.10)
$$\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}.$$

As the integrand in this example is nonnegative, the question of integrability does not arise. $\qquad \blacksquare$

Example 18.4. It is possible by means of Fubini's theorem to identify the limit in (17.4). First,

$$\int_0^t e^{-ux} \sin x\,dx = \frac{1}{1+u^2} \left[1 - e^{-ut}(u \sin t + \cos t)\right],$$

as follows by differentiation with respect to t. Since

$$\int_0^t \left[\int_0^{\infty} |e^{-ux} \sin x|\,du\right] dx = \int_0^t |\sin x| \cdot x^{-1}\,dx \leq t < \infty,$$

Fubini's theorem applies to the integration of $e^{-ux}\sin x$ over $(0,t) \times (0,\infty)$:

$$\int_0^t \frac{\sin x}{x}\,dx = \int_0^t \sin x \left[\int_0^\infty e^{-ux}\,du \right] dx$$

$$= \int_0^\infty \left[\int_0^t e^{-ux} \sin x\,dx \right] du$$

$$= \int_0^\infty \frac{du}{1+u^2} - \int_0^\infty \frac{e^{-ut}}{1+u^2}\,(u \sin t + \cos t)\,du.$$

The next-to-last integral is $\pi/2$ (see (17.9)), and a change of variable $ut = s$ shows that the final integral goes to 0 as $t \to \infty$. Therefore,

$$(18.11) \qquad\qquad \lim_{t \to \infty} \int_0^t \frac{\sin x}{x}\,dx = \frac{\pi}{2}. \qquad\qquad\blacksquare$$

Integration by Parts

Let F and G be two nondecreasing, right-continuous functions on an interval $[a, b]$, and let μ and ν be the corresponding measures:

$$\mu(x, y] = F(y) - F(x), \quad \nu(x, y] = G(y) - G(x), \qquad a \le x \le y \le b.$$

In accordance with the convention (17.22) write $dF(x)$ and $dG(x)$ in place of $\mu(dx)$ and $\nu(dx)$.

Theorem 18.4. *If F and G have no common points of discontinuity in $(a, b]$, then*

$$(18.12) \qquad \int_{(a,b\,]} G(x)\,dF(x)$$

$$= F(b)G(b) - F(a)G(a) - \int_{(a,b\,]} F(x)\,dG(x).$$

In brief: $\int G\,dF = \Delta FG - \int F\,dG$. This is one version of the partial integration formula.

PROOF. Note first that replacing $F(x)$ by $F(x) - C$ leaves (18.12) unchanged—it merely adds and subtracts $C\nu(a, b]$ on the right. Hence (take $C = F(a)$) it is no restriction to assume that $F(x) = \mu(a, x]$ and no restriction to assume that $G(x) = \nu(a, x]$. If $\pi = \mu \times \nu$ is product measure in the plane,

then by Fubini's theorem,

(18.13) $\pi[(x, y) : a < y \leq x \leq b]$

$$= \int_{(a, b]} \nu(a, x]\mu(dx) = \int_{(a, b]} G(x) \, dF(x)$$

and

(18.14) $\pi[(x, y) : a < x \leq y \leq b]$

$$= \int_{(a, b]} \mu(a, y]\nu(dy) = \int_{(a, b]} F(y) \, dG(y).$$

The two sets on the left have as their union the square $S = (a, b] \times (a, b]$. The diagonal of S has π-measure

$$\pi[(x, y) : a < x = y \leq b] = \int_{(a, b]} \nu\{x\}\mu(dx) = 0$$

because of the assumption that μ and ν share no points of positive measure. Thus the left sides of (18.13) and (18.14) add to $\pi(S) = \mu(a, b]\nu(a, b] = F(b)G(b)$. ∎

Suppose that ν has a density g with respect to Lebesgue measure and let $G(x) = c + \int_a^x g(t) \, dt$. Transform the right side of (18.12) by the formula (16.13) for integration with respect to a density; the result is

(18.15) $\int_{(a, b]} G(x) \, dF(x)$

$$= F(b)G(b) - F(a)G(a) - \int_a^b F(x)g(x) \, dx.$$

A consideration of positive and negative parts shows that this holds for any g integrable over $(a, b]$.

Suppose further that μ has a density f with respect to Lebesgue measure, and let $F(x) = c' + \int_a^x f(t) \, dt$. Then (18.15) further reduces to

(18.16) $\int_a^b G(x)f(x) \, dx = F(b)G(b) - F(a)G(a) - \int_a^b F(x)g(x) \, dx.$

Again, f can be any integrable function. This is the classical formula for integration by parts.

Under the appropriate integrability conditions, $(a, b]$ can be replaced by an unbounded interval.

Products of Higher Order

Suppose that $(X, \mathscr{X}, \mu)$, $(Y, \mathscr{Y}, \nu)$, and $(Z, \mathscr{Z}, \eta)$ are three σ-finite measure spaces. In the usual way, identify the products $X \times Y \times Z$ and $(X \times Y) \times Z$. Let $\mathscr{X} \times \mathscr{Y} \times \mathscr{Z}$ be the σ-field in $X \times Y \times Z$ generated by the $A \times B \times C$ with A, B, C in $\mathscr{X}, \mathscr{Y}, \mathscr{Z}$, respectively. For C in Z, let $\mathscr{E}_C$ be the class of $E \in \mathscr{X} \times \mathscr{Y}$ for which $E \times C \in \mathscr{X} \times \mathscr{Y} \times \mathscr{Z}$. Then $\mathscr{E}_C$ is a σ-field containing the measurable rectangles in $X \times Y$, and so $\mathscr{E}_C = \mathscr{X} \times \mathscr{Y}$. Therefore, $(\mathscr{X} \times \mathscr{Y}) \times \mathscr{Z} \subset \mathscr{X} \times \mathscr{Y} \times \mathscr{Z}$. But the reverse relation is obvious, and so $(\mathscr{X} \times \mathscr{Y}) \times \mathscr{Z} = \mathscr{X} \times \mathscr{Y} \times \mathscr{Z}$.

Define the product $\mu \times \nu \times \eta$ on $\mathscr{X} \times \mathscr{Y} \times \mathscr{Z}$ as $(\mu \times \nu) \times \eta$. It gives to $A \times B \times C$ the value $(\mu \times \nu)(A \times B) \cdot \eta(C) = \mu(A)\nu(B)\eta(C)$, and it is the only measure that does. The formulas (18.4) and (18.5) extend in the obvious way.

Products of four or more components can clearly be treated in the same way. This leads in particular to another construction of Lebesgue measure in $R^k = R^1 \times \cdots \times R^1$ (see Example 18.1) as the product $\lambda \times \cdots \times \lambda$ (k factors) on $\mathscr{R}^k = \mathscr{R}^1 \times \cdots \times \mathscr{R}^1$. Fubini's theorem of course gives a practical way to calculate volumes:

Example 18.5. Let V_k be the volume of the sphere of radius 1 in R^k; by Theorem 12.2, a sphere in R^k with radius r has volume $r^k V_k$. Let A be the unit sphere in R^k, let $B = [(x_1, x_2): x_1^2 + x_2^2 \le 1]$, and let $C(x_1, x_2) = [(x_3, \ldots, x_k): \sum_{i=3}^{k} \ldots x_i^2 \le 1 - x_1^2 - x_2^2]$. By Fubini's theorem,

$$V_k = \int_A dx_1 \cdots dx_k = \int_B dx_1 \, dx_2 \int_{C(x_1, x_2)} dx_3 \cdots dx_k$$

$$= \int_B dx_1 \, dx_2 \, V_{k-2}\left(1 - x_1^2 - x_2^2\right)^{(k-2)/2}$$

$$= V_{k-2} \iint_{\substack{0 < \theta < 2\pi \\ 0 < \rho < 1}} \left(1 - \rho^2\right)^{(k-2)/2} \rho \, d\rho \, d\theta$$

$$= \pi V_{k-2} \int_0^1 t^{(k-2)/2} \, dt = \frac{2\pi V_{k-2}}{k}.$$

If V_0 is taken as 1, this holds for $k = 2$ as well as for $k \ge 3$. Since $V_1 = 2$, it follows by induction that

$$V_{2i-1} = \frac{2(2\pi)^{i-1}}{1 \times 3 \times 5 \times \cdots \times (2i-1)}, \qquad V_{2i} = \frac{(2\pi)^i}{2 \times 4 \times \cdots \times (2i)}$$

for $i = 1, 2, \ldots$. Example 18.2 is a special case.   ∎

PROBLEMS

18.1. Show by Theorem 18.1 that if $A \times B$ is nonempty and lies in $\mathscr{X} \times \mathscr{Y}$, then $A \in \mathscr{X}$ and $B \in \mathscr{Y}$.

18.2. 2.9↑ Suppose that $X = Y$ is uncountable and $\mathscr{X} = \mathscr{Y}$ consists of the countable and the cocountable sets. Show that the diagonal $E = [(x, y): x = y]$ does not lie in $\mathscr{X} \times \mathscr{Y}$, even though $[y: (x, y) \in E] \in \mathscr{Y}$ and $[x: (x, y) \in E] \in \mathscr{X}$ for all x and y.

18.3. 10.5 18.1↑ Let $(X, \mathscr{X}, \mu) = (Y, \mathscr{Y}, \nu)$ be the completion of $(R^1, \mathscr{R}^1, \lambda)$. Show that $(X \times Y, \mathscr{X} \times \mathscr{Y}, \mu \times \nu)$ is not complete.

18.4. The assumption of σ-finiteness in Theorem 18.2 is essential: Let μ be Lebesgue measure on the line, let ν be counting measure on the line, and take $E = [(x, y): x = y]$. Then (18.1) and (18.2) do not agree.

18.5. Example 18.2 in effect connects π as the area of the unit disk D_1 with the π of trigonometry.

(a) A second way: Calculate $\lambda_2(D_1)$ directly by Fubini's theorem: $\lambda_2(D_1) = \int_{-1}^{1} 2(1 - x^2)^{1/2} dx$. Evaluate the integral by trigonometric substitution.

(b) A third way: Inscribe in the unit circle a regular polygon of n sides. Its interior consists of n congruent isosceles triangles with angle $2\pi/n$ at the apex; the area is $n \sin(\pi/n)\cos(\pi/n)$, which goes to π.

18.6. Suppose that f is nonnegative on a σ-finite measure space $(\Omega, \mathscr{F}, \mu)$. Show that

$$\int_\Omega f d\mu = (\mu \times \lambda)[(\omega, y) \in \Omega \times R^1 : 0 \leq y \leq f(\omega)].$$

Prove that the set on the right is measurable. This gives the "area under the curve." Given the existence of $\mu \times \lambda$ on $\Omega \times R^1$, one can use the right side of this equation as an alternative definition of the integral.

18.7. Reconsider Problem 12.12.

18.8. Suppose that $\nu[y: (x, y) \in E] = \nu[y: (x, y) \in F]$ for all x, and show that $(\mu \times \nu)(E) = (\mu \times \nu)(F)$. This is a general version of *Cavalieri's principle*.

18.9. (a) Suppose that μ is σ-finite, and prove the corollary to Theorem 16.7 by Fubini's theorem in the product of $(\Omega, \mathscr{F}, \mu)$ and $\{1, 2, \ldots\}$ with counting measure.

(b) Relate the series in Problem 17.7 to Fubini's theorem.

18.10. (a) Let $\mu = \nu$ be counting measure on $X = Y = \{1, 2, \ldots\}$. If

$$f(x, y) = \begin{cases} 2 - 2^{-x} & \text{if } x = y, \\ -2 + 2^{-x} & \text{if } x = y + 1, \\ 0 & \text{otherwise}, \end{cases}$$

then the iterated integrals exist but are unequal. Why does this not contradict Fubini's theorem?

(b) Show that $xy/(x^2 + y^2)^2$ is not integrable over the square $[(x, y) \mid |x|, |y| \leq 1]$ even though the iterated integrals exist and are equal

18.11. Exhibit a case in which (18.12) fails because F and G have a common point of discontinuity.

18.12. Prove (18.16) for the case in which all the functions are continuous by differentiating with respect to the upper limit of integration.

18.13. Prove for distribution functions F that $\int_{-\infty}^{\infty} (F(x + c) - F(x))\, dx = c$.

18.14. Prove for continuous distribution functions that $\int_{-\infty}^{\infty} F(x)\, dF(x) = \frac{1}{2}$.

18.15. Suppose that a number f_n is defined for each $n \geq n_0$ and put $F(x) = \sum_{n_0 \leq n \leq x} f_n$. Deduce from (18.15) that

$$(18.17) \qquad \sum_{n_0 \leq n \leq x} G(n) f_n = F(x) G(x) - \int_{n_0}^{x} F(t) g(t)\, dt$$

if $G(y) - G(x) = \int_x^y g(t)\, dt$, which will hold if G has continuous derivative g. First assume that the f_n are nonnegative.

18.16. ↑ Take $n_0 = 1$, $f_n = 1$, and $G(x) = 1/x$, and derive $\sum_{n \leq x} n^{-1} = \log x + \gamma + O(1/x)$, where $\gamma = 1 - \int_1^{\infty} (t - \lfloor t \rfloor) t^{-2}\, dt$ is Euler's constant.

18.17. 5.20 18.15↑ Use (18.17) and (5.51) to prove that there exists a constant c such that

$$(18.18) \qquad . \sum_{p \leq x} \frac{1}{p} = \log \log x + c + O\left(\frac{1}{\log x}\right).$$

18.18. Euler's *gamma function* is defined for positive t by $\Gamma(t) = \int_0^{\infty} x^{t-1} e^{-x}\, dx$.

(a) Prove that $\Gamma^{(k)}(t) = \int_0^{\infty} x^{t-1} (\log x)^k e^{-x}\, dx$.

(b) Show by partial integration that $\Gamma(t + 1) = t\Gamma(t)$ and hence that $\Gamma(n + 1) = n!$ for integral n.

(c) From (18.10) deduce $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

(d) Show that the unit sphere in R^k has volume (see Example 18.5)

$$(18.19) \qquad V_k = \frac{\pi^{k/2}}{\Gamma((k/2) + 1)}.$$

18.19. By partial integration prove that $\int_0^\infty (\sin x)/x)^2\,dx = \pi/2$ and $\int_{-\infty}^\infty (1 - \cos x)x^{-2}\,dx = \pi$.

18.20. Suppose that μ is a probability measure on $(X, \mathscr{X})$ and that, for each x in X, ν_x is a probability measure on $(Y, \mathscr{Y})$. Suppose further that, for each B in $\mathscr{Y}$, $\nu_x(B)$ is, as a function of x, measurable $\mathscr{X}$. Regard the $\mu(A)$ as initial probabilities and the $\nu_x(B)$ as transition probabilities:

   (a) Show that, if $E \in \mathscr{X} \times \mathscr{Y}$, then $\nu_x[y: (x, y) \in E]$ is measurable $\mathscr{X}$.

   (b) Show that $\pi(E) = \int_X \nu_x[y: (x, y) \in E]\mu(dx)$ defines a probability measure on $\mathscr{X} \times \mathscr{Y}$. If $\nu_x = \nu$ does not depend on x, this is just (18.1).

   (c) Show that if f is measurable $\mathscr{X} \times \mathscr{Y}$ and nonnegative, then $\int_Y f(x, y)\nu_x(dy)$ is measurable $\mathscr{X}$. Show further that

$$\int_{X \times Y} f(x, y)\pi(d(x, y)) = \int_X \left[\int_Y f(x, y)\nu_x(dy) \right]\mu(dx),$$

which extends Fubini's theorem (in the probability case). Consider also f's that may be negative.

   (d) Let $\nu(B) = \int_X \nu_x(B)\mu(dx)$. Show that $\pi(X \times B) = \nu(B)$ and

$$\int_Y f(y)\nu(dy) = \int_X \left[\int_Y f(y)\nu_x(dy) \right]\mu(dx).$$

SECTION 19. THE L^p SPACES*

Definitions

Fix a measure space $(\Omega, \mathscr{F}, \mu)$. For $1 \leq p < \infty$, let $L^p = L^p(\Omega, \mathscr{F}, \mu)$ be the class of (measurable) real functions f for which $|f|^p$ is integrable, and for f in L^p, write

$$(19.1) \qquad \|f\|_p = \left[\int |f|^p\,d\mu \right]^{1/p}.$$

There is a special definition for the case $p = \infty$: The *essential supremum* of f is

$$(19.2) \qquad \|f\|_\infty = \inf[\alpha: \mu[\omega: |f(\omega)| > \alpha] = 0];$$

take L^∞ to consist of those f for which this is finite. The spaces L^p have a geometric structure that can usefully guide the intuition. The basic facts are laid out in this section, together with two applications to theoretical statistics.

---

*The results in this section are used nowhere else in the book. The proofs require some elementary facts about metric spaces, vector spaces, and convex sets, and in one place the Radon–Nikodym theorem of Section 32 is used. As a matter of fact, it is possible to jump ahead and read Section 32 at this point, since it makes no use of Chapters 4 and 5

For $1 < p, q < \infty$, p and q are *conjugate* indices if $p^{-1} + q^{-1} = 1$; p and q are also conjugate if one is 1 and the other is ∞ (formally, $1^{-1} + \infty^{-1} = 1$). *Hölder's inequality* says that if p and q are conjugate and if $f \in L^p$ and $g \in L^q$, then fg is integrable and

$$(19.3) \qquad \left| \int fg \, d\mu \right| \le \int |fg| \, d\mu \le \|f\|_p \|g\|_q.$$

This is obvious if $p = 1$ and $q = \infty$. If $1 < p$, $q < \infty$ and μ is a probability measure, and if f and g are simple functions, (19.3) is (5.35). But the proof in Section 5 goes over without change to the general case.

Minkowski's inequality says that if $f, g \in L^p$ ($1 \le p \le \infty$), then $f + g \in L^p$ and

$$(19.4) \qquad \|f + g\|_p \le \|f\|_p + \|g\|_p.$$

This is clear if $p = 1$ or $p = \infty$. Suppose that $1 < p < \infty$. Since $|f + g| \le 2(|f|^p + |g|^p)^{1/p}$, $f + g$ does lie in L^p. Let q be conjugate to p. Since $p - 1 = p/q$, Hölder's inequality gives

$$\|f + g\|_p^p = \int |f + g|^p \, d\mu$$

$$\le \int |f| \cdot |f + g|^{p/q} \, d\mu + \int |g| \cdot |f + g|^{p/q} \, d\mu$$

$$\le \|f\|_p \cdot \left\| |f + g|^{p/q} \right\|_q + \|g\|_p \cdot \left\| |f + g|^{p/q} \right\|_q$$

$$= \left(\|f\|_p + \|g\|_p \right) \left[\int |f + g|^p \, d\mu \right]^{1/q}$$

$$= \left(\|f\|_p + \|g\|_p \right) \|f + g\|_p^{p/q}.$$

Since $p - p/q = 1$, (19.4) follows.[†]

If α is real and $f \in L^p$, then obviously $\alpha f \in L^p$ and

$$(19.5) \qquad \|\alpha f\|_p = |\alpha| \cdot \|f\|_p.$$

Define a metric on L^p by $d_p(f, g) = \|f - g\|_p$. Minkowski's inequality gives the triangle inequality for d_p, and d_p is certainly symmetric. Further, $d_p(f, g) = 0$ if and only if $|f - g|^p$ integrates to 0, that is, $f = g$ almost everywhere. To make L^p a metric space, identify functions that are equal almost everywhere.

[†]The Hölder and Minkowski inequalities can also be proved by convexity arguments, see Problems 5.9 and 5.10.

If $\|f - f_n\|_p \to 0$ and $p < \infty$, so that $\int |f - f_n|^p \, d\mu \to 0$, then f_n is said to converge to f *in the mean of order p*.

If $f = f'$ and $g = g'$ almost everywhere, then $f + g = f' + g'$ almost everywhere, and for real α, $\alpha f = \alpha f'$ almost everywhere. In L^p, f and f' become the same function, and similarly for the pairs g and g', $f + g$ and $f' + g'$, and αf and $\alpha f'$. This means that addition and scalar multiplication are unambiguously defined in L^p, which is thus a real vector space. It is a *normed vector space* in the sense that it is equipped with a *norm* $\| \cdot \|_p$ satisfying (19.4) and (19.5).

Completeness and Separability

A normed vector space is a *Banach* space if it is complete under the corresponding metric. According to the *Riesz–Fischer theorem*, this is true of L^p:

Theorem 19.1. *The space L^p is complete.*

PROOF. It is enough to show that every fundamental sequence contains a convergent subsequence. Suppose first that $p < \infty$. Assume that $\|f_m - f_n\|_p \to 0$ as $m, n \to \infty$, and choose an increasing sequence $\{n_k\}$ so that $\|f_m - f_n\|_p^p \le 2^{-(p+1)k}$ for $m, n \ge n_k$. Since $\int |f_m - f_n|^p \, d\mu \ge \alpha^p \mu[|f_m - f_n| \ge \alpha]$ (this is just a general version of Markov's inequality (5.31)), $\mu[|f_n - f_m| \ge 2^{-k}] \le 2^{pk}\|f_m - f_n\|_p^p \le 2^{-k}$ for $m, n \ge n_k$. Therefore, $\sum_l \mu[|f_{n_{k+1}} - f_{n_k}| \ge 2^{-k}] < \infty$, and it follows by the first Borel–Cantelli lemma (which works for arbitrary measures) that, outside a set of μ-measure 0, $\sum_k |f_{n_{k+1}} - f_{n_k}|$ converges. But then f_{n_k} converges to some f almost everywhere, and by Fatou's lemma, $\int |f - f_{n_k}|^p \, d\mu \le \liminf_i \int |f_{n_i} - f_{n_k}|^p \, d\mu \le 2^{-k}$. Therefore, $f \in L^p$ and $\|f - f_{n_k}\|_p \to 0$, as required.

If $p = \infty$, choose $\{n_k\}$ so that $\|f_m - f_n\|_\infty \le 2^{-k}$ for $m, n \ge n_k$. Since $|f_{n_{k+1}} - f_{n_k}| \le 2^{-k}$ almost everywhere, f_{n_k} converges to some f, and $|f - f_{n_k}| \le 2^{-k}$ almost everywhere. Again, $\|f - f_{n_k}\|_p \to 0$. ∎

The next theorem has to do with separability.

Theorem 19.2. (i) *Let U be the set of simple functions $\sum_{i=1}^m \alpha_i I_{B_i}$ for α_i and $\mu(B_i)$ finite. For $1 \le p \le \infty$, U is dense in L^p.*

(ii) *If μ is σ-finite and $\mathscr{F}$ is countably generated, and if $p < \infty$, then L^p is separable.*

PROOF. *Proof of* (i). Suppose first that $p < \infty$. For $f \in L^p$, choose (Theorem 13.5) simple functions f_n such that $f_n \to f$ and $|f_n| \le |f|$. Then $f_n \in L^p$, and by the dominated convergence theorem, $\int |f - f_n|^p \, d\mu \to 0$. Therefore,

$\|f - f_n\|_p < \epsilon$ for some n; but each f_n is in U. As for the case $p = \infty$, if $n2^n > \|f\|_\infty$, then the f_n defined by (13.6) satisfies $\|f - f_n\|_\infty \le 2^{-n}$ ($< \epsilon$ for large n).

Proof of (ii). Suppose that $\mathscr{F}$ is generated by a countable class $\mathscr{C}$ and that Ω is covered by a countable class $\mathscr{D}$ of $\mathscr{F}$-sets of finite measure. Let $E_1, E_2, \ldots$ be an enumeration of $\mathscr{C} \cup \mathscr{D}$; let $\mathscr{P}_n$ ($n \ge 1$) be the partition consisting of the sets of the form $F_1 \cap \cdots \cap F_n$, where F_i is E_i or E_i^c; and let $\mathscr{F}_n$ be the field of unions of the sets in $\mathscr{P}_n$. Then $\mathscr{F}_0 = \bigcup_{n=1}^{\infty} \mathscr{F}_n$ is a countable field that generates $\mathscr{F}$, and μ is σ-finite on $\mathscr{F}_0$. Let V be the set of simple functions $\sum_{i=1}^{m} \alpha_i I_{A_i}$ for α_i rational, $A_i \in \mathscr{F}_0$, and $\mu(A_i) < \infty$.

Let $g = \sum_{i=1}^{m} \alpha_i I_{B_i}$ be the element of U constructed in the proof of part (i). Then $\|f - g\|_p < \epsilon$, the α_i are rational by (13.6), and any α_i that vanish can be suppressed. By Theorem 11.4(ii), there exist sets A_i in $\mathscr{F}_0$ such that $\mu(B_i \triangle A_i) < (\epsilon/m|\alpha_i|)^p$, provided $p < \infty$, and then $h = \sum_{i=1}^{m} \alpha_i I_{A_i}$ lies in V and $\|f - h\|_p < 2\epsilon$. But V is countable.[†]   ∎

Conjugate Spaces

A *linear functional* on L^p is a real-valued function γ such that

$$(19.6) \qquad \gamma(\alpha f + \alpha' f') = \alpha \gamma(f) + \alpha' \gamma(f').$$

The functional is *bounded* if there is a finite M such that

$$(19.7) \qquad |\gamma(f)| \le M \cdot \|f\|_p$$

for all f in L^p. A bounded linear functional is uniformly continuous on L^p because $\|f - f'\|_p < \epsilon/M$ implies $|\gamma(f) - \gamma(f')| < \epsilon$ (if $M > 0$; and $M = 0$ implies $\gamma(f) \equiv 0$). The *norm* $\|\gamma\|$ of γ is the smallest M that works in (19.7): $\|\gamma\| = \sup[|\gamma(f)|/\|f\|_p : f \ne 0]$.

Suppose p and q are conjugate indices and $g \in L^q$. By Hölder's inequality,

$$(19.8) \qquad \gamma_g(f) = \int fg \, d\mu$$

is defined for $f \in L^p$ and satisfies (19.7) if $M \ge \|g\|_q$; and γ_g is obviously linear. According to the *Riesz representation theorem*, this is the most general bounded linear functional in the case $p < \infty$:

Theorem 19.3. *Suppose that μ is σ-finite, that $1 \le p < \infty$, and that q is conjugate to p. Every bounded linear functional on L^p has the form (19.8) for some $g \in L^q$; further,*

$$(19.9) \qquad \|\gamma_g\| = \|g\|_q,$$

and g is unique up to a set of μ-measure 0.

[†]Part (ii) definitely requires $p < \infty$; see Problem 19.2.

The space of bounded linear functionals on L^p is called the *dual space*, or the *conjugate space*, and the theorem identifies L^q as the dual of L^p. Note that the theorem does not cover the case $p = \infty$.[†]

PROOF. *Case I: μ finite.* For A in $\mathcal{F}$, define $\varphi(A) = \gamma(I_A)$. The linearity of γ implies that φ is finitely additive. For the M of (19.7), $|\varphi(A)| \le M \cdot \|I_A\|_p = M \cdot |\mu(A)|^{1/p}$. If $A = \bigcup_n A_n$, where the A_n are disjoint, then $\varphi(A) = \sum_{n=1}^N \varphi(A_n) + \varphi(\bigcup_{n>N} A_n)$, and since $|\varphi(\bigcup_{n>N} A_n)| \le M\mu^{1/p}(\bigcup_{n>N} A_n) \to 0$, it follows that φ is an additive set function in the sense of (32.1).

The Jordan decomposition (32.2) represents φ as the difference of two finite measures φ^+ and φ^- with disjoint supports A^+ and A^-. If $\mu(A) = 0$, then $\varphi^+(A) = \varphi(A \cap A^+) \le M\mu^{1/p}(A) = 0$. Thus φ^+ is absolutely continuous with respect to μ and by the Radon-Nikodym theorem (p. 422) has an integrable density g^+: $\varphi^+(A) = \int_A g^+ \, d\mu$. Together with the same result for φ^-, this shows that there is an integrable g such that $\gamma(I_A) = \varphi(A) = \int_A g \, d\mu = \int I_A g \, d\mu$. Thus $\gamma(f) = \int fg \, d\mu$ for simple functions f in L^p.

Assume that this g lies in L^q, and define γ_g by the equation (19.8). Then γ and γ_g are bounded linear functionals that agree for simple functions; since the latter are dense (Theorem 19.2(i)), it follows by the continuity of γ and γ_g that they agree on all of L^p. It is therefore enough (in the case of finite μ) to prove $g \in L^q$. It will also be shown that $\|g\|_q$ is at most the M of (19.7); since $\|g\|_q$ does work as a bound in (19.7), (19.9) will follow. If $\gamma_g(f) \equiv 0$, (19.9) will imply that $g = 0$ almost everywhere, and for the general γ it will follow further that two functions g satisfying $\gamma_g(f) \equiv \gamma(f)$ must agree almost everywhere.

Assume that $1 < p, q < \infty$. Let g_n be simple functions such that $0 \le g_n \uparrow |g|^q$, and take $h_n = g_n^{1/p} \operatorname{sgn} g$. Then $h_n g = g_n^{1/p}|g| \ge g_n^{1/p} g_n^{1/q} = g_n$, and since h_n is simple, it follows that $\int g_n \, d\mu \le \int h_n g \, d\mu = \gamma_g(h_n) = \gamma(h_n) \le M \cdot \|h_n\|_p = M[\int g_n \, d\mu]^{1/p}$. Since $1 - 1/p = 1/q$, this gives $[\int g_n \, d\mu]^{1/q} \le M$. Now the monotone convergence theorem gives $g \in L^p$ and even $\|g\|_q \le M$.

Assume that $p = 1$ and $q = \infty$. In this case, $|\int fg \, d\mu| = |\gamma_g(f)| = |\gamma(f)| \le M \cdot \|f\|_1$ for simple functions f in L^1. Take $f = \operatorname{sgn} g \cdot I_{[|g| \ge \alpha]}$. Then $\alpha\mu[|g| \ge \alpha] \le \int I_{[|g| \ge \alpha]} \cdot |g| \, d\mu = \int fg \, d\mu \le M \cdot \|f\|_1 = M\mu[|g| \ge \alpha]$. If $\alpha > M$, this inequality gives $\mu[|g| \ge \alpha] = 0$; therefore $\|g\|_\infty = \|g\|_q \le M$ and $g \in L^\infty = L^q$.

Case II: μ σ-finite. Let A_n be sets such that $A_n \uparrow \Omega$ and $\mu(A_n) < \infty$. If $\mu_n(A) = \mu(A \cap A_n)$, then $|\gamma(fI_{A_n})| \le M \cdot \|fI_{A_n}\|_p = M \cdot [\int |f|^p d\mu_n]^{1/p}$ for $f \in L^p(fI_{A_n} \in L^p(\mu) \subset L^p(\mu_n))$. By the finite case, A_n supports a g_n in L^q such that $\gamma(fI_{A_n}) = \int fI_{A_n} g_n \, d\mu$ for $f \in L^p$, and $\|g_n\|_q \le M$. Because of uniqueness, g_{n+1} can be taken to agree with g_n on A_n ($L^p(\mu_{n+1}) \subset L^p(\mu_n)$). There is therefore a function g on Ω such that $g = g_n$ on A_n and $\|I_{A_n} g\|_q \le M$. It follows that $\|g\|_q \le M$ and $g \in L^q$. By the dominated convergence theorem and the continuity of γ, $f \in L^p$ implies $\int fg \, d\mu = \lim_n \int fI_{A_n} g \, d\mu = \lim_n \gamma(fI_{A_n}) = \gamma(f)$. Uniqueness follows as before. ∎

Weak Compactness

For $f \in L^p$ and $g \in L^q$, where p and q are conjugate, write

(19.10)
$$(f, g) = \int fg \, d\mu.$$

For fixed f in L^p, this defines a bounded linear functional on L^q; for fixed g in L^q, it defines a bounded linear functional on L^p. By Hölder's inequality,

(19.11)
$$|(f, g)| \le \|f\|_p \|g\|_q.$$

Suppose that f and f_n are elements of L^p. If $(f, g) = \lim_n (f_n, g)$ for each g in L^q, then f_n *converges weakly* to f. If $\|f - f_n\|_p \to 0$, then certainly f_n converges weakly to f, although the converse is false.[†]

The next theorem says in effect that if $p > 1$, then the unit ball $B_1^p = \{f \in L^p : \|f\|_p \le 1\}$ is compact in the topology of weak convergence.

Theorem 19.4. *Suppose that μ is σ-finite and $\mathscr{F}$ is countably generated. If $1 < p \le \infty$, then every sequence in B_1^p contains a subsequence converging weakly to an element of B_1^p.*

Suppose of elements f_n, f, and f' of L^p that f_n converges weakly both to f and to f'. Since, by hypothesis, μ is σ-finite and $p > 1$, Theorem 19.3 applies if the p and q there are interchanged. And now, since $(f, g) = (f', g)$ for all g in L^q, it follows by uniqueness that $f = f'$. Therefore, weak limits are unique under the present hypothesis. The assumption $p > 1$ is essential.[‡]

PROOF. Let q be conjugate to p ($1 \le q < \infty$). By Theorem 19.2(ii), L^q contains a countable, dense set G. Add to G all finite, rational linear combinations of its elements; it is still countable. Suppose that $\{f_n\} \subset B_1^p$.

By (19.11), $|(f_n, g)| \le \|g\|_q$ for $g \in L^q$. Since $\{(f_n, g)\}$ is bounded, it is possible by the diagonal method [A14] to pass to a subsequence of $\{f_n\}$ along which, for each of the countably many g in G, the limit $\lim_n (f_n, g) = \gamma(g)$ exists and $|\gamma(g)| \le \|g\|_q$. For $g, g' \in G$, $|\gamma(g) - \gamma(g')| = \lim_n |(f_n, g - g')| \le \|g - g'\|_q$. Therefore, γ is uniformly continuous on G and so has a unique continuous extension to all of L^q. For $g, g' \in G$, $\gamma(g + g') = \lim_n (f_n, g + g') = \gamma(g) + \gamma(g')$; by continuity, this extends to all of L^q. For $g \in G$ and α rational, $\gamma(\alpha g) = \alpha \lim_n (f_n, g) = \alpha \gamma(g)$; by continuity, this extends to all real α and all g in L^q: γ is a linear functional on L^q. Finally, $|\gamma(g)| \le \|g\|_q$ extends from G to L^q by continuity, and γ is bounded in the sense of (19.7).

[†]Problem 19.4
[‡]Problem 19.5.

By the Riesz representation theorem ($1 \leq q < \infty$), there is an f in L^p (the space adjoint to L^q) such that $\gamma(g) = (f, g)$ for all g. Since γ has norm at most 1, (19.9) implies that $\|f\|_p \leq 1$: f lies in B_1^p.

Now $(f, g) = \lim_n (f_n, g)$ for g in G. Suppose that g' is an arbitrary element of L^q, and choose g in G so that $\|g' - g\|_q < \epsilon$. Then

$$|(f, g') - (f_n, g')|$$

$$\leq |(f, g') - (f, g)| + |(f, g) - (f_n, g)| + |(f_n, g) - (f_n, g')|$$

$$\leq \|f\|_p \|g' - g\|_q + |(f, g) - (f_n, g)| + \|f_n\|_p \|g - g'\|_q$$

$$\leq 2\epsilon + |(f, g) - (f_n, g)|.$$

Since $g \in G$, the last term here goes to 0, and hence $\lim_n (f_n, g') = (f, g')$ for all g' in L^q. Therefore, f_n converges weakly to f. $\blacksquare$

Some Decision Theory

The weak compactness of the unit ball in L^∞ has interesting implications for statistical decision theory. Suppose that μ is σ-finite and $\mathcal{F}$ is countably generated. Let $f_1, \ldots, f_k$ be probability densities with respect to μ—nonnegative and integrating to 1. Imagine that, for some i, ω is drawn from Ω according to the probability measure $P_i(A) = \int_A f_i \, d\mu$. The statistical problem is to decide, on the basis of an observed ω, which f_i is the right one.

Assume that if the right density is f_i, then a statistician choosing f_j incurs a nonnegative *loss* $L(j|i)$. A *decision rule* is a vector function $\delta(\omega) = (\delta_1(\omega), \ldots, \delta_k(\omega))$, where the $\delta_i(\omega)$ are nonnegative and add to 1: the statistician, observing ω, selects f_i with probability $\delta_i(\omega)$. If, for each ω, $\delta_i(\omega)$ is 1 for one i and 0 for the others, δ is a *nonrandomized* rule; otherwise, it is a *randomized* rule. Let D be the set of all rules. The problem is to choose, in some more or less rational way that connects up with the losses $L(j|i)$, a rule δ from D.

The *risk* corresponding to δ and f_i is

$$R_i(\delta) = \int \left[\sum_j \delta_j(\omega) L(j|i) \right] f_i(\omega) \mu(d\omega),$$

which can be interpreted as the loss a statistician using δ can expect if f_i is the right density. The *risk point* for δ is $R(\delta) = (R_1(\delta), \ldots, R_k(\delta))$. If $R_i(\delta') \leq R_i(\delta)$ for all i and $R_i(\delta') < R_i(\delta)$ for some i—that is, if the point $R(\delta')$ is "southwest" of $R(\delta)$—then of course δ' is taken as being *better* than δ. There is in general no rule better than all the others. (Different rules can have the same risk point, but they are then indistinguishable as regards the decision problem.)

The *risk set* is the collection S of all the risk points; it is a bounded set in the first orthant of R^k. To avoid trivialities, assume that S does not contain the origin (as would happen for example if the $L(j|i)$ were all 0).

Suppose that δ and δ' are elements of D, and λ and λ' are nonnegative and add to 1. If $\delta_i''(\omega) = \lambda \delta_i(\omega) + \lambda' \delta_i'(\omega)$, then δ'' is in D and $R(\delta'') = \lambda R(\delta) + \lambda' R(\delta')$. Therefore, S is a convex set.

Lying much deeper is the fact that S is compact. Given points $x^{(n)}$ in S, choose rules $\delta^{(n)}$ such that $R(\delta^{(n)}) = x^{(n)}$. Now $\delta_j^{(n)}(\cdot)$ is an element of L^∞, in fact of B_1^∞, and so by Theorem 19.4 there is a subsequence along which, for each $j = 1, \ldots, k$, $\delta_j^{(n)}$

converges weakly to a function δ_j in B_1^∞. If $\mu(A) < \infty$, then $\int \delta_j I_A \, d\mu = \lim_n \int \delta_j^{(n)} I_A \, d\mu \geq 0$ and $\int (1 - \sum_j \delta_j) I_A \, d\mu = \lim_n \int (1 - \sum_j \delta_j^{(n)}) I_A \, d\mu = 0$, so that $\delta_j \geq 0$ and $\sum_j \delta_j = 1$ almost everywhere on A. Since μ is σ-finite, the δ_j can be altered on a set of μ-measure 0 in such a way as to ensure that $\delta = (\delta_1, \ldots, \delta_k)$ is an element of D. But, along the subsequence, $x^{(n)} \to R(\delta)$. Therefore: *The risk set is compact and convex*.

The rest is geometry. For x in R^k, let Q_x be the set of x' such that $0 \leq x_i' \leq x_i$ for all i. If $x = R(\delta)$ and $x' = R(\delta')$, then δ' is *better than* δ if and only if $x' \in Q_x$ and $x' \neq x$. A rule δ is *admissible* if there exists no δ' better than δ; it makes no sense to use a rule that is not admissible. Geometrically, admissibility means that, for $x = R(\delta)$, $S \cap Q_x$ consists of x alone.
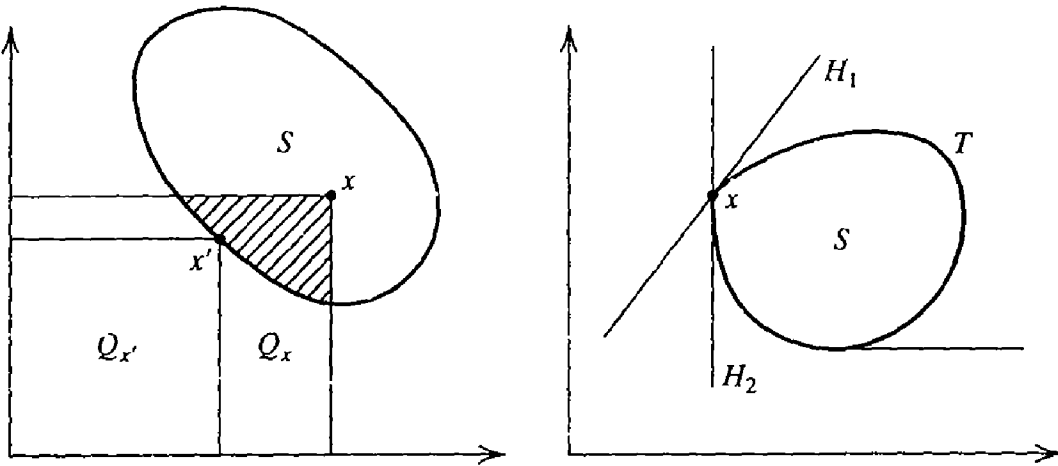
Let $x = R(\delta)$ be given, and suppose that δ is not admissible. Since $S \cap Q_x$ is compact, it contains a point x' nearest the origin (x' unique, since $S \cap Q_x$ is convex as well as compact); let δ' be a corresponding rule: $x' = R(\delta')$. Since δ is not admissible, $x' \neq x$, and δ' is better than δ. If $S \cap Q_{x'}$ contained a point distinct from x', it would be a point of $S \cap Q_x$ nearer the origin than x', which is impossible. This means that $Q_{x'}$ contains no point of S other than x' itself, which means in turn that δ' is admissible. Therefore, if δ is not itself admissible, there is a δ' that is admissible and is better than δ. This is expressed by saying that the class of admissible rules is *complete*.

Let $p = (p_1, \ldots, p_k)$ be a probability vector, and view p_i as an a priori probability that f_i is the correct density. A rule δ has *Bayes risk* $R(p, \delta) = \sum_i p_i R_i(\delta)$ with respect to p. This is a kind of compound risk: f_i is correct with probability p_i, and the statistician chooses f_j with probability $\delta_j(\omega)$. A *Bayes rule* is one that minimizes the Bayes risk for a given p. In this case, take $\alpha = R(p, \delta)$ and consider the hyperplane

$$(19.12) \qquad H = \left[z: \sum_i p_i z_i = \alpha \right]$$

and the half space

$$(19.13) \qquad H^+ = \left[z: \sum_i p_i z_i \geq \alpha \right].$$

Then $x = R(\delta)$ lies on H, and S is contained in H^+: x is on the boundary of S, and

H is a supporting hyperplane. If $p_i > 0$ for all i, then Q_x meets S only at x, and so δ is admissible.

Suppose now that δ is admissible, so that $x = R(\delta)$ is the only point in $S \cap Q_x$ and x lies on the boundary of S. The problem is to show that δ is a Bayes rule, which means finding a supporting hyperplane (19.12) corresponding to a probability vector p. Let T consist of those y for which Q_y meets S. Then T is convex: given a convex combination $y'' = \lambda y + \lambda' y'$ of points in T, choose in S points z and z' southwest of y and y', respectively, and note that $z'' = \lambda z + \lambda' z'$ lies in S and is southwest of y''. Since S meets Q_x only in the point x, the same is true of T, so that x is a boundary point of T as well as of S. Let (19.12) ($p \neq 0$) be a supporting hyperplane through x: $x \in H$ and $T \subset H^+$. If $p_{i_0} < 0$, take $z_{i_0} = x_{i_0} + 1$ and take $z_i = x_i$ for the other i; then z lies in T but not in H^+, a contradiction. (The right-hand figure shows the role of T: the planes H_1 and H_2 both support S, but only H_2 supports T and only H_2 corresponds to a probability vector.) Thus $p_i \geq 0$ for all i, and since $\Sigma_i p_i = 1$ can be arranged by normalization, δ is indeed a Bayes rule. Therefore *The admissible rules are Bayes rules, and they form a complete class.*

The Space L^2

The space L^2 is special because $p = 2$ is its own conjugate index. If $f, g \in L^2$, the *inner product* $(f, g) = \int fg \, d\mu$ is well defined, and by (19.11)—write $\|f\|$ in place of $\|f\|_2$—$|(f, g)| \leq \|f\| \cdot \|g\|$. This is the *Schwarz* (or Cauchy–Schwarz) inequality. If one of f and g is fixed, (f, g) is a bounded (hence continuous) linear functional in the other. Further, $(f, g) = (g, f)$, the norm is given by $\|f\|^2 = (f, f)$, and L^2 is complete under the metric $d(f, g) = \|f - g\|$. A *Hilbert space* is a vector space on which is defined an inner product having all these properties.

The Hilbert space L^2 is quite like Euclidean space. If $(f, g) = 0$, then f and g are *orthogonal*, and orthogonality is like perpendicularity. If $f_1, \ldots, f_n$ are orthogonal (in pairs), then by linearity, $(\Sigma_i f_i, \Sigma_j f_j) = \Sigma_i \Sigma_j (f_i, f_j) = \Sigma_i (f_i, f_i)$: $\|\Sigma_i f_i\|^2 = \Sigma_i \|f_i\|^2$. This is a version of the Pythagorean theorem. If f and g are orthogonal, write $f \perp g$. For every f, $f \perp 0$.

Suppose now that μ is σ-finite and $\mathcal{F}$ is countably generated, so that L^2 is separable as a metric space. The construction that follows gives a sequence (finite or infinite) $\varphi_1, \varphi_2, \ldots$ that is *orthonormal* in the sense that $\|\varphi_n\| = 1$ for all n and $(\varphi_m, \varphi_n) = 0$ for $m \neq n$, and is *complete* in the sense that $(f, \varphi_n) = 0$ for all n implies $f = 0$—so that the orthonormal system cannot be enlarged. Start with a sequence $f_1, f_2, \ldots$ that is dense in L^2. Define $g_1, g_2, \ldots$ inductively: Let $g_1 = f_1$. Suppose that $g_1, \ldots, g_n$ have been defined and are orthogonal. Define $g_{n+1} = f_{n+1} - \Sigma_{i=1}^n \alpha_{ni} g_i$, where α_{ni} is $(f_{n+1}, g_i)/\|g_i\|^2$ if $g_i \neq 0$ and is arbitrary if $g_i = 0$. Then g_{n+1} is orthogonal to $g_1, \ldots, g_n$, and f_{n+1} is a linear combination of $g_1, \ldots, g_{n+1}$. This, the Gram–Schmidt method, gives an orthogonal sequence $g_1, g_2, \ldots$ with the property that the finite linear combinations of the g_n include all the f_n and are therefore dense in L^2. If $g_n \neq 0$, take $\varphi_n = g_n/\|g_n\|$; if $g_n = 0$, discard it from the sequence. Then $\varphi_1, \varphi_2, \ldots$ is orthonormal, and the finite linear combinations of the φ_n are still dense. It can happen that all but finitely many of the g_n are 0, in which

case there are only finitely many of the φ_n. In what follows it is assumed that $\varphi_1, \varphi_2, \ldots$ is an infinite sequence; the finite case is analogous and somewhat simpler.

Suppose that f is orthogonal to all the φ_n. If a_i are arbitrary scalars, then $f, a_1\varphi_1, \ldots, a_n\varphi_n$ is an orthogonal set, and by the Pythagorean property, $\|f - \sum_{i=1}^{n} a_i\varphi_i\|^2 = \|f\|^2 + \sum_{i=1}^{n} a_i^2 \geq \|f\|^2$. If $\|f\| > 0$, then f cannot be approximated by finite linear combinations of the φ_n, a contradiction: $\varphi_1, \varphi_2, \ldots$ is a complete orthonormal system.

Consider now a sequence $a_1, a_2, \ldots$ of scalars for which $\sum_{i=1}^{\infty} a_i^2$ converges. If $s_n = \sum_{i=1}^{n} a_i\varphi_i$, then the Pythagorean theorem gives $\|s_n - s_m\|^2 = \sum_{m < i \leq n} a_i^2$. Since the scalar series converges, $\{s_n\}$ is fundamental and therefore by Theorem 19.1 converges to some g in L^2. Thus $g = \lim_n \sum_{i=1}^{n} a_i\varphi_i$, which it is natural to express as $g = \sum_{i=1}^{\infty} a_i\varphi_i$. The series (that is to say, the sequence of partial sums) converges to g in the mean of order 2 (not almost everywhere). By the following argument, every element of L^2 has a unique representation in this form.

The *Fourier coefficients* of f with respect to $\{\varphi_i\}$ are the inner products $a_i = (f, \varphi_i)$. For each n, $0 \leq \|f - \sum_{i=1}^{n} a_i\varphi_i\|^2 = \|f\|^2 - 2\sum_i a_i(f, \varphi_i) + \sum_{ij} a_i a_j(\varphi_i, \varphi_j) = \|f\|^2 - \sum_{i=1}^{n} a_i^2$, and hence, n being arbitrary, $\sum_{i=1}^{\infty} a_i^2 \leq \|f\|^2$. By the argument above, the series $\sum_{i=1}^{\infty} a_i\varphi_i$ therefore converges. By linearity, $(f - \sum_{i=1}^{n} a_i\varphi_i, \varphi_j) = 0$ for $n \geq j$, and by continuity, $(f - \sum_{i=1}^{\infty} a_i\varphi_i, \varphi_j) = 0$. Therefore, $f - \sum_{i=1}^{\infty} a_i\varphi_i$ is orthogonal to each φ_j and by completeness must be 0:

$$(19.14) \qquad\qquad f = \sum_{i=1}^{\infty} (f, \varphi_i)\varphi_i.$$

This is the *Fourier representation* of f. It is unique because if $f = \sum_{i=1}^{\infty} a_i\varphi_i$ is 0 ($\sum a_i^2 < \infty$), then $a_j = (f, \varphi_j) = 0$. Because of (19.14), $\{\varphi_n\}$ is also called an orthonormal *basis* for L^2.

A subset M of L^2 is a *subspace* if it is closed both algebraically ($f, f' \in M$ implies $\alpha f + \alpha' f' \in M$) and topologically ($f_n \in M$, $f_n \to f$ implies $f \in M$). If L^2 is separable, then so is the subspace M, and the construction above carries over: M contains an orthonormal system $\{\varphi_n\}$ that is *complete in M*, in the sense that $f = 0$ if $(f, \varphi_n) = 0$ for all n *and if* $f \in M$. And each f in M has the unique Fourier representation (19.14). Even if f does not lie in M, $\sum_{i=1}^{\infty} (f, \varphi_i)^2$ converges, so that $\sum_{i=1}^{\infty} (f, \varphi_i)\varphi_i$ is well defined.

This leads to a powerful idea, that of *orthogonal projection* onto M. For an orthonormal basis $\{\varphi_i\}$ of M, define $P_M f = \sum_{i=1}^{\infty} (f, \varphi_i)\varphi_i$ for all f in L^2 (not just for f in M). Clearly, $P_M f \in M$. Further, $f - \sum_{i=1}^{n} (f, \varphi_i)\varphi_i \perp \varphi_j$ for $n \geq j$ by linearity, so that $f - P_M f \perp \varphi_j$ by continuity. But if $f - P_M f$ is orthogonal to each φ_j, then, again by linearity and continuity, it is orthogonal to the general element $\sum_{j=1}^{\infty} b_j\varphi_j$ of M. Therefore, $P_M f \in M$ and $f - P_M f \perp M$. The map $f \to P_M f$ is the orthogonal projection on M.

The fundamental properties of P_M are these:

(i) $g \in M$ and $f - g \perp M$ together imply $g = P_M f$;

(ii) $f \in M$ implies $P_M f = f$;

(iii) $g \in M$ implies $\|f - g\| \geq \|f - P_M f\|$;

(iv) $P_M(\alpha f + \alpha' f') = \alpha P_M f + \alpha' P_M f'$.

Property (i) says that $P_M f$ is uniquely determined by the two conditions $P_M f \in M$ and $f - P_M \perp M$. To prove it, suppose that $g, g' \in M$, $f - g \perp M$, and $f - g' \perp M$. Then $g - g' \in M$ and $g - g' \perp M$, so that $g - g'$ is orthogonal to itself and hence $\|g - g'\|^2 = 0$: $g = g'$. Thus the mapping P_M is independent of the particular basis $\{\varphi_i\}$; it is determined by M alone.

Clearly, (ii) follows from (i); it implies that P_M is idempotent in the sense that $P_M^2 f = P_M f$. As for (iii), if g lies in M, so does $P_M f - g$, so that, by the Pythagorean relation, $\|f - g\|^2 = \|f - P_M f\|^2 + \|P_M f - g\|^2 \geq \|f - P_M f\|^2$; the inequality is strict if $g \neq P_M f$. Thus $P_M f$ is the unique point of M lying nearest to f. Property (iv), linearity, follows from (i).

An Estimation Problem

First, the technical setting: Let $(\Omega, \mathscr{F}, \mu)$ and $(\Theta, \mathscr{E}, \pi)$ be a σ-finite space and a probability space, and assume that $\mathscr{F}$ and $\mathscr{E}$ are countably generated. Let $f_\theta(\omega)$ be a nonnegative function on $\Theta \times \Omega$, measurable $\mathscr{E} \times \mathscr{F}$, and assume that $\int_\Omega f_\theta(\omega) \mu(d\omega) = 1$ for each $\theta \in \Theta$. For some unknown value of θ, ω is drawn from Ω according to the probabilities $P_\theta(A) = \int_A f_\theta(\omega) \mu(d\omega)$, and the statistical problem is to estimate the value of $g(\theta)$, where g is a real function on Θ. The statistician knows the functions $f(\cdot)$ and $g(\cdot)$, as well as the value of ω; it is the value of θ that is unknown.

For an example, take Ω to be the line, $f(\omega)$ a function known to the statistician, and $f_\theta(\omega) = \alpha f(\alpha\omega + \beta)$, where $\theta = (\alpha, \beta)$ specifies unknown scale and location parameters; the problem is to estimate $g(\theta) = \alpha$, say. Or, more simply, as in the exponential case (14.7), take $f_\theta(\omega) = \alpha f(\alpha\omega)$, where $\theta = g(\theta) = \alpha$.

An *estimator* of $g(\theta)$ is a function $t(\omega)$. It is *unbiased* if

$$(19.15) \qquad \int_\Omega t(\omega) f_\theta(\omega) \mu(d\omega) = g(\theta)$$

for all θ in Θ (assume the integral exists); this condition means that the estimate is on target in an average sense. A natural loss function is $(t(\omega) - g(\theta))^2$, and if f_θ is the correct density, the *risk* is taken to be $\int_\Omega (t(\omega) - g(\theta))^2 f_\theta(\omega) \mu(d\theta)$.

If the probability measure π is regarded as an a priori distribution for the unknown θ, the *Bayes risk* of t is

$$(19.16) \qquad R(\pi, t) = \int_\Theta \int_\Omega (t(\omega) - g(\theta))^2 f_\theta(\omega) \mu(d\omega) \pi(d\theta);$$

this integral, assumed finite, can be viewed as a joint integral or as an iterated integral (Fubini's theorem). And now t_0 is a *Bayes estimator* of g with respect to π if it minimizes $R(\pi, t)$ over t. This is analogous to the Bayes rules discussed earlier. The

following simple projection argument shows that, except in trivial cases, no Bayes estimator is unbaised [†]

Let Q be the probability measure on $\mathscr{E} \times \mathscr{F}$ having density $f_\theta(\omega)$ with respect to $\pi \times \mu$, and let L^2 be the space of square-integrable functions on $(\Theta \times \Omega, \mathscr{E} \times \mathscr{F}, Q)$. Then Q is finite and $\mathscr{E} \times \mathscr{F}$ is countably generated. Recall that an element of L^2 is an equivalence class of functions that are equal almost everywhere with respect to Q. Let G be the class of elements of L^2 containing a function of the form $\bar{g}(\theta, \omega) = g(\omega)$ —functions of θ alone Then G is a subspace. (That G is algebraically closed is clear; if $f_n \in G$ and $\|f_n - f\| \to 0$, then—see the proof of Theorem 19.1—some subsequence converges to f outside a set of Q-measure 0, and it follows easily that $f \in G$.) Similarly, let T be the subspace of functions of ω alone: $\bar{t}(\theta, \omega) = t(\omega)$. Consider only functions g and their estimators t for which the corresponding $\bar{g}$ and $\bar{t}$ are in L^2.

Suppose now that t_0 is both an unbiased estimator of g_0 and a Bayes estimator of g_0 with respect to π. By (19 16) for g_0, $R(\pi, t) = \|\bar{t} - \bar{g}_0\|^2$, and since t_0 is a Bayes estimator of g_0, it follows that $\|\bar{t}_0 - \bar{g}_0\|^2 \le \|\bar{t} - \bar{g}_0\|^2$ for all $\bar{t}$ in T. This means that $\bar{t}_0$ is the orthogonal projection of $\bar{g}_0$ on the subspace T and hence that $\bar{g}_0 - \bar{t}_0 \perp \bar{t}_0$. On the other hand, from the assumption that t_0 is an unbiased estimator of g_0, it follows that, for every $\bar{g}(\theta, \omega) = g(\theta)$ in G,

$$(\bar{t}_0 - \bar{g}_0, \bar{g}) = \int_\Theta \int_\Omega (t_0(\omega) - g_0(\theta)) g(\theta) f_\theta(\omega) \mu(d\omega) \pi(d\theta)$$

$$= \int_\Theta g(\theta) \left[\int_\Omega (t_0(\omega) - g_0(\theta)) f_\theta(\omega) \mu(d\omega) \right] \pi(d\theta) = 0.$$

This means that $\bar{t}_0 - \bar{g}_0 \perp G$: $\bar{g}_0$ is the orthogonal projection of $\bar{t}_0$ on the subspace G But $\bar{g}_0 - \bar{t}_0 \perp \bar{t}_0$ and $\bar{t}_0 - \bar{g}_0 \perp \bar{g}_0$ together imply that $\bar{t}_0 - \bar{g}_0$ is orthogonal to itself: $\bar{t}_0 = \bar{g}_0$. Therefore, $t_0(\omega) = \bar{t}_0(\theta, \omega) = \bar{g}_0(\theta, \omega) = g_0(\theta)$ for (θ, ω) outside a set of Q-measure 0.

This implies that t_0 and g_0 are essentially constant. Suppose for simplicity that $f_\theta(\omega) > 0$ for all (θ, ω), so that (Theorem 15.2) $(\pi \times \mu)[(\theta, \omega): t_0(\omega) \ne g_0(\theta)] = 0$. By Fubini's theorem, there is a θ such that, if $a = g_0(\theta)$, then $\mu[\omega: t_0(\omega) \ne a] = 0$; and there is an ω such that, if $b = t_0(\omega)$, then $\pi[\theta: g_0(\theta) \ne b] = 0$. It follows that, for (θ, ω) outside a set of $(\pi \times \mu)$-measure 0, $t_0(\omega)$ and $g_0(\theta)$ have the common value $a = b$: $\pi[\theta: g_0(\theta) = a] = 1$ and $P_\theta[\omega: t_0(\omega) = a] = 1$ for all θ in Θ.

PROBLEMS

19.1. Suppose that $\mu(\Omega) < \infty$ and $f \in L^\infty$. Show that $\|f\|_p \uparrow \|f\|_\infty$.

19.2. (a) Show that $L^\infty((0, 1], \mathscr{B}, \lambda)$ is not separable.

   (b) Show that $L^p((0, 1], \mathscr{B}, \mu)$ is not separable if μ is counting measure (μ is not σ-finite).

   (c) Show that $L^p(\Omega, \mathscr{F}, P)$ is not separable if (Theorem 36.2) there is on the space an independent stochastic process $[X_t: 0 \le t \le 1]$ such that X_t takes the values ± 1 with probability $\frac{1}{2}$ each ($\mathscr{F}$ is not countably generated).

[†]This is interesting because of the close connection between Bayes rules and admissibility; see BERGER, pp. 546 ff.

19.3. Show that Theorem 19.3 fails for $L^\infty((0,1], \mathcal{B}, \lambda)$. *Hint:* Take $\gamma(f)$ to be a Banach limit of $n\int_0^{1/n} f(x)\,dx$.

19.4. Consider weak convergence in $L^p((0,1], \mathcal{B}, \lambda)$.

(a) For the case $p = \infty$, find functions f_n and f such that f_n goes weakly to f but $\|f - f_n\|_p$ does not go to 0.

(b) Do the same for $p = 2$.

19.5. Show that the unit ball in $L^1((0,1], \mathcal{B}, \lambda)$ is not weakly compact.

19.6. Show that a Bayes rule corresponding to $p = (p_1, \ .\ ,p_k)$ may not be admissible if $p_i = 0$ for some i But there will be a better Bayes rule that *is* admissible

19.7. *The Neyman–Pearson lemma.* Suppose f_1 and f_2 are rival densities and $L(j|i)$ is 0 or 1 as $j = i$ or $j \neq i$, so that $R_i(\delta)$ is the probability of choosing the opposite density when f_i is the right one. Suppose of δ that $\delta_2(\omega) = 1$ if $f_2(\omega) > tf_1(\omega)$ and $\delta_2(\omega) = 0$ if $f_2(\omega) < tf_1(\omega)$, where $t > 0$ Show that δ is admissible: For any rule δ', $\int \delta'_2 f_1\,d\mu < \int \delta_2 f_1\,d\mu$ implies $\int \delta'_1 f_2\,d\mu > \int \delta_1 f_2\,d\mu$. *Hint.* $\int (\delta_2 - \delta'_2)(f_2 - tf_1)\,d\mu \geq 0$, since the integrand is nonnegative.

19.8. The classical orthonormal basis for $L^2[0, 2\pi]$ with Lebesgue measure is the trigonometric system

$$(19.17) \quad (2\pi)^{-1}, \qquad \pi^{-1/2}\sin nx, \qquad \pi^{-1/2}\cos nx, \qquad n = 1, 2, ..$$

Prove orthonormality. *Hint:* Express the sines and cosines in terms of $e^{inx} \pm e^{-inx}$, multiply out the products, and use the fact that $\int_0^{2\pi} e^{imx}\,dx$ is 2π or 0 as $m = 0$ or $m \neq 0$. (For the completeness of the trigonometric system, see Problem 26.26.)

19.9. Drop the assumption that L^2 is separable. Order by inclusion the orthonormal systems in L^2, and let (Zorn's lemma) $\Phi = [\varphi_\gamma : \gamma \in \Gamma]$ be maximal.

(a) Show that $\Gamma_f = [\gamma : (f, \varphi_\gamma) \neq 0]$ is countable. *Hint·* Use $\sum_{i=1}^n (f, \varphi_\gamma)^2 \leq \|f\|^2$ and the argument for Theorem 10.2(iv).

(b) Let $Pf = \sum_{\gamma \in \Gamma_f} (f, \varphi_\gamma)\varphi_\gamma$. Show that $f - Pf \perp \Phi$ and hence (maximality) $f = Pf$. Thus Φ is an orthonormal basis.

(c) Show that Φ is countable if and only if L^2 is separable.

(d) Now take Φ to be a maximal orthonormal system in a subspace M, and define $P_M f = \sum_{\gamma \in \Gamma_f} (f, \varphi_\gamma)\varphi_\gamma$. Show that $P_M f \in M$ and $f - P_M f \perp \Phi$, that $g = P_M g$ if $g \in M$, and that $f - P_M f \perp M$. This defines the general orthogonal projection.

Random Variables and Expected Values

SECTION 20. RANDOM VARIABLES AND DISTRIBUTIONS

This section and the next cover random variables and the machinery for dealing with them—expected values, distributions, moment generating functions, independence, convolution.

Random Variables and Vectors

A *random variable* on a probability space $(\Omega, \mathscr{F}, P)$ is a real-valued function $X = X(\omega)$ measurable $\mathscr{F}$. Sections 5 through 9 dealt with random variables of a special kind, namely simple random variables, those with finite range. All concepts and facts concerning real measurable functions carry over to random variables; any changes are matters of viewpoint, notation, and terminology only.

The positive and negative parts X^+ and X^- of X are defined as in (15.4) and (15.5). Theorem 13.5 also applies: Define

$$(20.1) \qquad \psi_n(x) = \begin{cases} (k-1)2^{-n} & \text{if } (k-1)2^{-n} \leq x < k2^{-n}, \\ & \qquad\qquad 1 \leq k \leq n2^n, \\ n & \text{if } x \geq n. \end{cases}$$

If X is nonnegative and $X_n = \psi_n(X)$, then $0 \leq X_n \uparrow X$. If X is not necessarily nonnegative, define

$$(20.2) \qquad X_n = \begin{cases} \psi_n(X) & \text{if } X \geq 0, \\ -\psi_n(-X) & \text{if } X \leq 0. \end{cases}$$

(This is the same as (13.6).) Then $0 \leq X_n(\omega) \uparrow X(\omega)$ if $X(\omega) \geq 0$ and $0 \geq$

$X_n(\omega) \downarrow X(\omega)$ if $X(\omega) \le 0$; and $|X_n(\omega)| \uparrow |X(\omega)|$ for every ω. The random variable X_n is in each case simple.

A *random vector* is a mapping from Ω to R^k that is measurable $\mathscr{F}$. Any mapping from Ω to R^k must have the form $\omega \to X(\omega) = (X_1(\omega), \ldots, X_k(\omega))$, where each $X_i(\omega)$ is real; as shown in Section 13 (see (13.2)), X is measurable if and only if each X_i is. Thus a random vector is simply a k-tuple $X = (X_1, \ldots, X_k)$ of random variables.

Subfields

If $\mathscr{G}$ is a σ-field for which $\mathscr{G} \subset \mathscr{F}$, a k-dimensional random vector X is of course measurable $\mathscr{G}$ if $[\omega: X(\omega) \in H] \in \mathscr{G}$ for every H in $\mathscr{R}^k$. The σ-field $\sigma(X)$ generated by X is the smallest σ-field with respect to which it is measurable. The σ-field generated by a collection of random vectors is the smallest σ-field with respect to which each one is measurable.

As explained in Sections 4 and 5, a sub-σ-field corresponds to partial information about ω. The information contained in $\sigma(X) = \sigma(X_1, \ldots, X_k)$ consists of the k numbers $X_1(\omega), \ldots, X_k(\omega)$.[†] The following theorem is the analogue of Theorem 5.1, but there are technical complications in its proof.

Theorem 20.1. *Let $X = (X_1, \ldots, X_k)$ be a random vector.*

(i) *The σ-field $\sigma(X) = \sigma(X_1, \ldots, X_k)$ consists exactly of the sets $[X \in H]$ for $H \in \mathscr{R}^k$.*

(ii) *In order that a random variable Y be measurable $\sigma(X) = \sigma(X_1, \ldots, X_k)$ it is necessary and sufficient that there exist a measurable map $f: R^k \to R^1$ such that $Y(\omega) = f(X_1(\omega), \ldots, X_k(\omega))$ for all ω.*

PROOF. The class $\mathscr{G}$ of sets of the form $[X \in H]$ for $H \in \mathscr{R}^k$ is a σ-field. Since X is measurable $\sigma(X)$, $\mathscr{G} \subset \sigma(X)$. Since X is measurable $\mathscr{G}$, $\sigma(X) \subset \mathscr{G}$. Hence part (i).

Measurability of f in part (ii) refers of course to measurability $\mathscr{R}^k/\mathscr{R}^1$. The sufficiency is easy: if such an f exists, Theorem 13.1(ii) implies that Y is measurable $\sigma(X)$.

To prove necessity,[‡] suppose at first that Y is a simple random variable, and let $y_1, \ldots, y_m$ be its different possible values. Since $A_i = [\omega: Y(\omega) = y_i]$ lies in $\sigma(X)$, it must by part (i) have the form $[\omega: X(\omega) \in H_i]$ for some H_i in $\mathscr{R}^k$. Put $f = \sum_i y_i I_{H_i}$; certainly f is measurable. Since the A_i are disjoint, no $X(\omega)$ can lie in more than one H_i (even though the latter need not be disjoint), and hence $f(X(\omega)) = Y(\omega)$.

[†] The partition defined by (4.16) consists of the sets $[\omega. \; X(\omega) = x]$ for $x \in R^k$.

[‡] For a general version of this argument, see Problem 13.3.

To treat the general case, consider simple random variables Y_n such that $Y_n(\omega) \to Y(\omega)$ for each ω. For each n, there is a measurable function f_n: $R^k \to R^1$ such that $Y_n(\omega) = f_n(X(\omega))$ for all ω. Let M be the set of x in R^k for which $\{f_n(x)\}$ converges; by Theorem 13.4(iii), M lies in $\mathscr{R}^k$. Let $f(x) = \lim_n f_n(x)$ for x in M, and let $f(x) = 0$ for x in $R^k - M$. Since $f = \lim_n f_n I_M$ and $f_n I_M$ is measurable, f is measurable by Theorem 13.4(ii). For each ω, $Y(\omega) = \lim_n f_n(X(\omega))$; this implies in the first place that $X(\omega)$ lies in M and in the second place that $Y(\omega) = \lim_n f_n(X(\omega)) = f(X(\omega))$. ∎

Distributions

The distribution or law of a random variable X was in Section 14 defined as the probability measure on the line given by $\mu = PX^{-1}$ (see (13.7)), or

$$(20.3) \qquad \mu(A) = P[X \in A], \qquad A \in \mathscr{R}^1.$$

The distribution function of X was defined by

$$(20.4) \qquad F(x) = \mu(-\infty, x] = P[X \le x]$$

for real x. The left-hand limit of F satisfies

$$(20.5) \qquad \begin{aligned} F(x-) &= \mu(-\infty, x) = P[X < x], \\ F(x) - F(x-) &= \mu\{x\} = P[X = x], \end{aligned}$$

and F has at most countably many discontinuities. Further, F is nondecreasing and right-continuous, and $\lim_{x \to -\infty} F(x) = 0$, $\lim_{x \to \infty} F(x) = 1$. By Theorem 14.1, for each F with these properties there exists on some probability space a random variable having F as its distribution function.

A support for μ is a Borel set S for which $\mu(S) = 1$. A random variable, its distribution, and its distribution function are *discrete* if μ has a countable support $S = \{x_1, x_2, \dots\}$. In this case μ is completely determined by the values $\mu\{x_1\}, \mu\{x_2\}, \dots$.

A familiar discrete distribution is the *binomial*:

$$(20.6) \qquad P[X = r] = \mu\{r\} = \binom{n}{r} p^r (1-p)^{n-r}, \qquad r = 0, 1, \dots, n.$$

There are many random variables, on many spaces, with this distribution: If $\{X_k\}$ is an independent sequence such that $P[X_k = 1] = p$ and $P[X_k = 0] = 1 - p$ (see Theorem 5.3), then X could be $\sum_{i=1}^n X_i$, or $\sum_{i=9}^{8+n} X_i$, or the sum of any n of the X_i. Or Ω could be $\{0, 1, \dots, n\}$ if $\mathscr{F}$ consists of all subsets, $P\{r\} = \mu\{r\}$, $r = 0, 1, \dots, n$, and $X(r) \equiv r$. Or again the space and random variable could be those given by the construction in either of the two proofs of Theorem 14.1. These examples show that, although the distribution of a

random variable X contains all the information about the probabilistic behavior of X itself, it contains beyond this no further information about the underlying probability space $(\Omega, \mathscr{F}, P)$ or about the interaction of X with other random variables on the space.

Another common discrete distribution is the *Poisson* distribution with parameter $\lambda > 0$:

$$(20.7) \qquad P[X = r] = \mu\{r\} = e^{-\lambda} \frac{\lambda^r}{r!}, \quad r = 0, 1, \ldots .$$

A *constant* c can be regarded as a discrete random variable with $X(\omega) \equiv c$. In this case $P[X = c] = \mu\{c\} = 1$. For an artificial discrete example, let $\{x_1, x_2, \ldots\}$ be an enumeration of the rationals, and put

$$(20.8) \qquad \mu\{x_r\} = 2^{-r};$$

the point of the example is that the support need not be contained in a lattice.

A random variable and its distribution have *density* f with respect to Lebesgue measure if f is a nonnegative Borel function on R^1 and

$$(20.9) \qquad P[X \in A] = \mu(A) = \int_A f(x)\, dx, \qquad A \in \mathscr{R}^1.$$

In other words, the requirement is that μ have density f with respect to Lebesgue measure λ in the sense of (16.11). The density is assumed to be with respect to λ if no other measure is specified.

Taking $A = R^1$ in (20.9) shows that f must integrate to 1. Note that f is determined only to within a set of Lebesgue measure 0: if $f = g$ except on a set of Lebesgue measure 0, then g can also serve as a density for X and μ.

It follows by Theorem 3.3 that (20.9) holds for every Borel set A if it holds for every interval—that is, if

$$F(b) - F(a) = \int_a^b f(x)\, dx$$

holds for every a and b. Note that F need not differentiate to f everywhere (see (20.13), for example); all that is required is that f integrate properly—that (20.9) hold. On the other hand, if F does differentiate to f and f is continuous, it follows by the fundamental theorem of calculus that f is indeed a density for F.[†]

[†] The general question of the relation between differentiation and integration is taken up in Section 31

For the *exponential distribution* with parameter $\alpha > 0$, the density is

$$(20.10) \qquad f(x) = \begin{cases} 0 & \text{if } x < 0, \\ \alpha e^{-\alpha x} & \text{if } x \geq 0. \end{cases}$$

The corresponding distribution function

$$(20.11) \qquad F(x) = \begin{cases} 0 & \text{if } x \leq 0, \\ 1 - e^{-\alpha x} & \text{if } \geq 0 \end{cases}$$

was studied in Section 14.

For the *normal distribution* with parameters m and σ, $\sigma > 0$,

$$(20.12) \qquad f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{(x-m)^2}{2\sigma^2} \right], \qquad -\infty < x < \infty;$$

a change of variable together with (18.10) shows that f does integrate to 1. For the *standard* normal distribution, $m = 0$ and $\sigma = 1$.

For the *uniform* distribution over an interval $(a, b]$,

$$(20.13) \qquad f(x) = \begin{cases} \dfrac{1}{b-a} & \text{if } a < x \leq b, \\ 0 & \text{otherwise.} \end{cases}$$

The distribution function F is useful if it has a simple expression, as in (20.11). It is ordinarily simpler to describe μ by means of a density $f(x)$ or discrete probabilities $\mu\{x_r\}$.

If F comes from a density, it is continuous. In the discrete case, F increases in jumps; the example (20.8), in which the points of discontinuity are dense, shows that it may nonetheless be very irregular. There exist distributions that are not discrete but are not continuous either. An example is $\mu(A) = \frac{1}{2}\mu_1(A) + \frac{1}{2}\mu_2(A)$ for μ_1 discrete and μ_2 coming from a density; such mixed cases arise, but they are few. Section 31 has examples of a more interesting kind, namely functions F that are continuous but do not come from any density. These are the functions singular in the sense of Lebesgue; the $Q(x)$ describing bold play in gambling (see (7.33)) turns out to be one of them. See Example 31.1.

If X has distribution μ and g is a real function of a real variable,

$$(20.14) \qquad P[g(X) \in A] = P[X \in g^{-1}A] = \mu(g^{-1}A).$$

Thus the distribution of $g(X)$ is μg^{-1} in the notation (13.7).

In the case where there is a density, f and F are related by

$$(20.15) \qquad F(x) = \int_{-\infty}^{x} f(t)\, dt.$$

Hence f at its continuity points must be the derivative of F. As noted above, if F has a continuous derivative, this derivative can serve as the density f. Suppose that f is continuous and g is increasing, and let $T = g^{-1}$. The distribution function of $g(X)$ is $P[g(X) \leq x] = P[X \leq T(x)] = F(T(x))$. If T is differentiable, this differentiates to $f(T(x))T'(x)$, which is therefore the density for $g(X)$. If g is decreasing, on the other hand, then $P[g(X) < x] = P[X > T(x)] = 1 - F(T(x))$, and the derivative is equal to $-f(T(x))T'(x) = f(T(x))|T'(x)|$. In either case, $g(X)$ has density

$$(20.16) \qquad \frac{d}{dx} P[g(X) \leq x] = f(T(x))|T'(x)|.$$

If X has the normal density (20.12) and $a > 0$, (20.16) shows that $aX + b$ has the normal density with parameters $am + b$ and $a\sigma$. Finding the density of $g(X)$ from first principles, as in the argument leading to (20.16), often works even if g is many-to-one:

Example 20.1. If X has the standard normal distribution, then

$$P[X^2 \leq x] = \frac{1}{\sqrt{2\pi}} \int_{-\sqrt{x}}^{\sqrt{x}} e^{-t^2/2}\, dt = \frac{2}{\sqrt{2\pi}} \int_{0}^{\sqrt{x}} e^{-t^2/2}\, dt$$

for $x > 0$. Hence X^2 has density

$$f(x) = \begin{cases} 0 & \text{if } x \leq 0, \\ \dfrac{1}{\sqrt{2\pi}} x^{-1/2} e^{-x/2} & \text{if } x > 0. \end{cases} \qquad \blacksquare$$

Multidimensional Distributions

For a k-dimensional random vector $X = (X_1, \ldots, X_k)$, the distribution μ (a probability measure on $\mathscr{R}^k$) and the distribution function F (a real function on R^k) are defined by

$$(20.17) \qquad \begin{aligned} \mu(A) &= P[(X_1, \ldots, X_k) \in A], \qquad A \in \mathscr{R}^k, \\ F(x_1, \ldots, x_k) &= P[X_1 \leq x_1, \ldots, X_k \leq x_k] = \mu(S_x), \end{aligned}$$

where $S_x = [y: y_i \leq x_i,\ i = 1, \ldots, k]$ consists of the points "southwest" of x.

Often μ and F are called the *joint* distribution and *joint* distribution function of $X_1, \ldots, X_k$.

Now F is nondecreasing in each variable, and $\Delta_A F \geq 0$ for bounded rectangles A (see (12.12)). As h decreases to 0, the set

$$S_{x,h} = [\, y\colon y_i \leq x_i + h, i = 1, \ldots, k \,]$$

decreases to S_x, and therefore (Theorem 2.1(ii)) F is continuous from above in the sense that $\lim_{h \downarrow 0} F(x_1 + h, \ldots, x_k + h) = F(x_1, \ldots, x_k)$. Further, $F(x_1, \ldots, x_k) \to 0$ if $x_i \to -\infty$ for some i (the other coordinates held fixed), and $F(x_1, \ldots, x_k) \to 1$ if $x_i \to \infty$ for each i. For any F with these properties there is by Theorem 12.5 a unique probability measure μ on $\mathscr{R}^k$ such that $\mu(A) = \Delta_A F$ for bounded rectangles A, and $\mu(S_x) = F(x)$ for all x.

As h decreases to 0, $S_{x, -h}$ increases to the interior $S_x^\circ = [\, y\colon y_i < x_i, i = 1, \ldots, k \,]$ of S_x, and so

$$(20.18) \qquad \lim_{h \downarrow 0} F(x_1 - h, \ldots, x_k - h) = \mu(S_x^\circ).$$

Since F is nondecreasing in each variable, it is continuous at x if and only if it is continuous from below there in the sense that this last limit coincides with $F(x)$. Thus F is continuous at x if and only if $F(x) = \mu(S_x) = \mu(S_x^\circ)$, which holds if and only if the boundary $\partial S_x = S_x - S_x^\circ$ (the y-set where $y_i \leq x_i$ for all i and $y_i = x_i$ for some i) satisfies $\mu(\partial S_x) = 0$. If $k > 1$, F can have discontinuity points even if μ has no point masses: if μ corresponds to a uniform distribution of mass over the segment $B = [(x, 0)\colon 0 < x < 1]$ in the plane ($\mu(A) = \lambda[x\colon 0 < x < 1, (x, 0) \in A]$), then F is discontinuous at each point of B. This also shows that F can be discontinuous at uncountably many points. On the other hand, for fixed x the boundaries $\partial S_{x,h}$ are disjoint for different values of h, and so (Theorem 10.2(iv)) only countably many of them can have positive μ-measure. Thus x is the limit of points $(x_1 + h, \ldots, x_k + h)$ at which F is continuous: the continuity points of F are dense.

There is always a random vector having a given distribution and distribution function: Take $(\Omega, \mathscr{F}, P) = (R^k, \mathscr{R}^k, \mu)$ and $X(\omega) \equiv \omega$. This is the obvious extension of the construction in the first proof of Theorem 14.1.

The distribution may as for the line be discrete in the sense of having countable support. It may have density f with respect to k-dimensional Lebesgue measure: $\mu(A) = \int_A f(x)\,dx$. As in the case $k = 1$, the distribution μ is more fundamental than the distribution function F, and usually μ is described not by F but by a density or by discrete probabilities.

If X is a k-dimensional random vector and $g\colon R^k \to R^i$ is measurable, then $g(X)$ is an i-dimensional random vector; if the distribution of X is μ, the distribution of $g(X)$ is μg^{-1}, just as in the case $k = 1$—see (20.14). If $g_j\colon R^k \to R^1$ is defined by $g_j(x_1, \ldots, x_k) = x_j$, then $g_j(X)$ is X_j, and its distribution $\mu_j = \mu g_j^{-1}$ is given by $\mu_j(A) = \mu[(x_1, \ldots, x_k)\colon x_j \in A] = P[X_j \in A]$ for

$A \in \mathscr{R}^1$. The μ_j are the *marginal distributions* of μ. If μ has a density f in R^k, then μ_j has over the line the density

(20.19)

$$f_j(x) = \int_{R^{k-1}} f(x_1, \ldots, x_{j-1}, x, x_{j+1}, \ldots, x_k) \, dx_1 \cdots dx_{j-1} \, dx_{j+1} \cdots dx_k,$$

since by Fubini's theorem the right side integrated over A comes to $\mu[(x_1, \ldots, x_k): x_j \in A]$.

Now suppose that g is a one-to-one, continuously differentiable map of V onto U, where U and V are open sets in R^k. Let T be the inverse, and suppose its Jacobian $J(x)$ never vanishes. If X has a density f supported by V, then for $A \subset U$, $P[g(X) \in A] = P[X \in TA] = \int_{TA} f(y) \, dy$, and by (17.10), this equals $\int_A f(Tx)|J(x)| \, dx$. Therefore, $g(X)$ has density

(20.20)
$$d(x) = \begin{cases} f(Tx)|J(x)| & \text{for } x \in U, \\ 0 & \text{for } x \notin U. \end{cases}$$

This is the analogue of (20.16).

Example 20.2. Suppose that (X_1, X_2) has density

$$f(x_1, x_2) = (2\pi)^{-1} \exp\left[-\tfrac{1}{2}(x_1^2 + x_2^2) \right],$$

and let g be the transformation to polar coordinates. Then U, V, and T are as in Example 17.7. If R and Θ are the polar coordinates of (X_1, X_2), then $(R, \Theta) = g(X_1, X_2)$ has density $(2\pi)^{-1} \rho e^{-\rho^2/2}$ in V. By (20.19), R has density $\rho e^{-\rho^2/2}$ on $(0, \infty)$, and Θ is uniformly distributed over $(0, 2\pi)$. ∎

For the normal distribution in R^k, see Section 29.

Independence

Random variables $X_1, \ldots, X_k$ are defined to be independent if the σ-fields $\sigma(X_1), \ldots, \sigma(X_k)$ they generate are independent in the sense of Section 4. This concept for simple random variables was studied extensively in Chapter 1; the general case was touched on in Section 14. Since $\sigma(X_i)$ consists of the sets $[X_i \in H]$ for $H \in \mathscr{R}^1$, $X_1, \ldots, X_k$ are independent if and only if

(20.21) $$P[X_1 \in H_1, \ldots, X_k \in H_k] = P[X_1 \in H_1] \cdots P[X_k \in H_k]$$

for all linear Borel sets $H_1, \ldots, H_k$. The definition (4.10) of independence requires that (20.21) hold also if some of the events $[X_i \in H_i]$ are suppressed on each side, but this only means taking $H_i = R^1$.

Suppose that

$$(20.22) \qquad P[X_1 \le x_1, \ldots, X_k \le x_k] = P[X_1 \le x_1] \cdots P[X_k \le x_k]$$

for all real $x_1, \ldots, x_k$; it then also holds if some of the events $[X_i \le x_i]$ are suppressed on each side (let $x_i \to \infty$). Since the intervals $(-\infty, x]$ form a π-system generating $\mathscr{R}^1$, the sets $[X_i \le x]$ form a π-system generating $\sigma(X_i)$. Therefore, by Theorem 4.2, (20.22) implies that $X_1, \ldots, X_k$ are independent. If, for example, the X_i are integer-valued, it is enough that $P[X_1 = n_1, \ldots, X_k = n_k] = P[X_1 = n_1] \cdots P[X_k = n_k]$ for integral $n_1, \ldots, n_k$ (see (5.9)).

Let $(X_1, \ldots, X_k)$ have distribution μ and distribution function F, and let the X_i have distributions μ_i and distribution functions F_i (the marginals). By (20.21), $X_1, \ldots, X_k$ are independent if and only if μ is product measure in the sense of Section 18:

$$(20.23) \qquad \mu = \mu_1 \times \cdots \times \mu_k.$$

By (20.22), $X_1, \ldots, X_k$ are independent if and only if

$$(20.24) \qquad F(x_1, \ldots, x_k) = F_1(x_1) \cdots F_k(x_k).$$

Suppose that each μ_i has density f_i; by Fubini's theorem, $f_1(y_1) \cdots f_k(y_k)$ integrated over $(-\infty, x_1] \times \cdots \times (-\infty, x_k]$ is just $F_1(x_1) \cdots F_k(x_k)$, so that μ has density

$$(20.25) \qquad f(x) = f_1(x_1) \cdots f_k(x_k)$$

in the case of independence.

If $\mathscr{G}_1, \ldots, \mathscr{G}_k$ are independent σ-fields and X_i is measurable $\mathscr{G}_i$, $i = 1, \ldots, k$, then certainly $X_1, \ldots, X_k$ are independent.

If X_i is a d_i-dimensional random vector, $i = 1, \ldots, k$, then $X_1, \ldots, X_k$ are by definition independent if the σ-fields $\sigma(X_1), \ldots, \sigma(X_k)$ are independent. The theory is just as for random variables: $X_1, \ldots, X_k$ are independent if and only if (20.21) holds for $H_1 \in \mathscr{R}^{d_1}, \ldots, H_k \in \mathscr{R}^{d_k}$. Now $(X_1, \ldots, X_k)$ can be regarded as a random vector of dimension $d = \sum_{i=1}^k d_i$; if μ is its distribution in $R^d = R^{d_1} \times \cdots \times R^{d_k}$ and μ_i is the distribution of X_i in R^{d_i}, then, just as before, $X_1, \ldots, X_k$ are independent if and only if $\mu = \mu_1 \times \cdots \times \mu_k$. In none of this need the d_i components of a single X_i be themselves independent random variables.

An infinite collection of random variables or random vectors is by definition independent if each finite subcollection is. The argument following (5.10)

extends from collections of simple random variables to collections of random vectors:

Theorem 20.2. *Suppose that*

$$(20.26) \qquad \begin{matrix} X_{11} & X_{12} & \cdots \\ X_{21} & X_{22} & \cdots \\ \vdots & \vdots \end{matrix}$$

is an independent collection of random vectors. If $\mathscr{F}_i$ is the σ-field generated by the ith row, then $\mathscr{F}_1, \mathscr{F}_2, \ldots$ are independent.

PROOF. Let $\mathscr{A}_i$ consist of the finite intersections of sets of the form $[X_{ij} \in H]$ with H a Borel set in a space of the appropriate dimension, and apply Theorem 4.2. The σ-fields $\mathscr{F}_i = \sigma(\mathscr{A}_i)$, $i = 1, \ldots, n$, are independent for each n, and the result follows. ∎

Each row of (20.26) may be finite or infinite, and there may be finitely or infinitely many rows. As a matter of fact, rows may be uncountable and there may be uncountably many of them.

Suppose that X and Y are independent random vectors with distributions μ and ν in R^j and R^k. Then (X, Y) has distribution $\mu \times \nu$ in $R^j \times R^k = R^{j+k}$. Let x range over R^j and y over R^k. By Fubini's theorem,

$$(20.27) \quad (\mu \times \nu)(B) = \int_{R^j} \nu[y : (x, y) \in B] \mu(dx), \qquad B \in \mathscr{R}^{j+k}.$$

Replace B by $(A \times R^k) \cap B$, where $A \in \mathscr{R}^j$ and $B \in \mathscr{R}^{j+k}$. Then (20.27) reduces to

$$(20.28) \quad (\mu \times \nu)((A \times R^k) \cap B) = \int_A \nu[y : (x, y) \in B] \mu(dx),$$

$$A \in \mathscr{R}^j, \quad B \in \mathscr{R}^{j+k}.$$

If $B_x = [y : (x, y) \in B]$ is the x-section of B, so that $B_x \in \mathscr{R}^k$ (Theorem 18.1), then $P[(x, Y) \in B] = P[\omega : (x, Y(\omega)) \in B] = P[\omega : Y(\omega) \in B_x] = \nu(B_x)$. Expressing the formulas in terms of the random vectors themselves gives this result:

Theorem 20.3. *If X and Y are independent random vectors with distributions μ and ν in R^j and R^k, then*

$$(20.29) \quad P[(X, Y) \in B] = \int_{R^j} P[(x, Y) \in B] \mu(dx), \qquad B \in \mathscr{R}^{j+k},$$

and

(20.30) $P[X \in A, (X, Y) \in B] = \int_A P[(x, Y) \in B] \mu(dx),$

$$A \in \mathcal{R}^j, \quad B \in \mathcal{R}^{j+k}.$$

Example 20.3. Suppose that X and Y are independent exponentially distributed random variables. By (20.29), $P[Y/X \geq z] = \int_0^\infty P[Y/x \geq z] \alpha e^{-\alpha x} dx = \int_0^\infty e^{-\alpha xz} \alpha e^{-\alpha x} dx = (1 + z)^{-1}$. Thus Y/X has density $(1 + z)^{-2}$ for $z \geq 0$. Since $P[X \geq z_1, Y/X \geq z_2] = \int_{z_1}^\infty P[Y/x \geq z_2] \alpha e^{-\alpha x} dx$ by (20.30), the joint distribution of X and Y/X can be calculated as well.   ∎

The formulas (20.29) and (20.30) are constantly applied as in this example. There is no virtue in making an issue of each case, however, and the appeal to Theorem 20.3 is usually silent.

Example 20.4. Here is a more complicated argument of the same sort. Let $X_1, \ldots, X_n$ be independent random variables, each uniformly distributed over $[0, t]$. Let Y_k be the kth smallest among the X_i, so that $0 \leq Y_1 \leq \cdots \leq Y_n \leq t$. The X_i divide $[0, t]$ into $n + 1$ subintervals of lengths $Y_1, Y_2 - Y_1, \ldots, Y_n - Y_{n-1}, t - Y_n$; let M be the maximum of these lengths. Define $\psi_n(t, a) = P[M \leq a]$. The problem is to show that

(20.31) $\psi_n(t, a) = \sum_{k=0}^{n+1} (-1)^k \binom{n+1}{k} \left(1 - k\frac{a}{t}\right)^n_+,$

where $x_+ = (x + |x|)/2$ denotes positive part.

Separate consideration of the possibilities $0 \leq a \leq t/2$, $t/2 \leq a \leq t$, and $t \leq a$ disposes of the case $n = 1$. Suppose it is shown that the probability $\psi_n(t, a)$ satisfies the recursion

(20.32) $\psi_n(t, a) = n \int_0^a \psi_{n-1}(t - x, a) \left(\frac{t - x}{t}\right)^{n-1} \frac{dx}{t}.$

Now (as follows by an integration together with Pascal's identity for binomial coefficients) the right side of (20.31) satisfies this same recursion, and so it will follow by induction that (20.31) holds for all n.

In intuitive form, the argument for (20.32) is this: If $[M \leq a]$ is to hold, the smallest of the X_i must have some value x in $[0, a]$. If X_1 is the smallest of the X_i, then $X_2, \ldots, X_n$ must all lie in $[x, t]$ and divide it into subintervals of length at most a; the probability of this is $(1 - x/t)^{n-1} \psi_{n-1}(t - x, a)$, because $X_2, \ldots, X_n$ have probability $(1 - x/t)^{n-1}$ of all lying in $[x, t]$, and if they do, they are independent and uniformly distributed there. Now (20.32) results from integrating with respect to the density for X_1 and multiplying by n to allow for the fact that any of $X_1, \ldots, X_n$ may be the smallest.

To make this argument rigorous, apply (20.30) for $j = 1$ and $k = n - 1$. Let A be the interval $[0, a]$, and let B consist of the points $(x_1, \ldots, x_n)$ for which $0 \leq x_i \leq t$, x_1 is the minimum of $x_1, \ldots, x_n$, and $x_2, \ldots, x_n$ divide $[x_1, t]$ into subintervals of length at most a. Then $P[X_1 = \min X_i, M \leq a] = P[X_1 \in A, (X_1, \ldots, X_n) \in B]$. Take X_1 for

X and $(X_2, \ldots, X_n)$ for Y in (20.30). Since X_1 has density $1/t$,

$$(20.33) \qquad P[X_1 = \min X_i, M \le a] = \int_0^a P[(x, X_2, \ldots, X_n) \in B] \frac{dx}{t}.$$

If C is the event that $x \le X_i \le t$ for $2 \le i \le n$, then $P(C) = (1 - x/t)^{n-1}$. A simple calculation shows that $P[X_i - x \le s_i, 2 \le i \le n | C] = \prod_{i=2}^n (s_i/(t-x))$; in other words, given C, the random variables $X_2 - x, \ldots, X_n - x$ are conditionally independent and uniformly distributed over $[0, t - x]$. Now $X_2, \ldots, X_n$ are random variables on some probability space $(\Omega, \mathscr{F}, P)$; replacing P by $P(\cdot|C)$ shows that the integrand in (20.33) is the same as that in (20.32). The same argument holds with the index 1 replaced by any k $(1 \le k \le n)$, which gives (20.32). (The events $[X_k = \min X_i, Y \le a]$ are not disjoint, but any two intersect in a set of probability 0.) ∎

Sequences of Random Variables

Theorem 5.3 extends to general distributions μ_n.

Theorem 20.4. *If $\{\mu_n\}$ is a finite or infinite sequence of probability measures on $\mathscr{R}^1$, there exists on some probability space $(\Omega, \mathscr{F}, P)$ an independent sequence $\{X_n\}$ of random variables such that X_n has distribution μ_n.*

PROOF. By Theorem 5.3 there exists on some probability space an independent sequence $Z_1, Z_2, \ldots$ of random variables assuming the values 0 and 1 with probabilities $P[Z_n = 0] = P[Z_n = 1] = \frac{1}{2}$. As a matter of fact, Theorem 5.3 is not needed: take the space to be the unit interval and the $Z_n(\omega)$ to be the digits of the dyadic expansion of ω—the functions $d_n(\omega)$ of Sections and 1 and 4.

Relabel the countably many random variables Z_n so that they form a double array,

$$
\begin{array}{ccc}
Z_{11} & Z_{12} & \cdots \\
Z_{21} & Z_{22} & \cdots \\
\vdots & \vdots &
\end{array}
$$

All the Z_{nk} are independent. Put $U_n = \sum_{k=1}^\infty Z_{nk} 2^{-k}$ The series certainly converges, and U_n is a random variable by Theorem 13.4. Further, $U_1, U_2, \ldots$ is, by Theorem 20.2, an independent sequence.

Now $P[Z_{ni} = z_i, 1 \le i \le k] = 2^{-k}$ for each sequence $z_1, \ldots, z_k$ of 0's and 1's; hence the 2^k possible values $j2^{-k}$, $0 \le j < 2^k$, of $S_{nk} = \sum_{i=1}^k Z_{ni} 2^{-i}$ all have probability 2^{-k}. If $0 \le x < 1$, the number of the $j2^{-k}$ that lie in $[0, x]$ is $\lfloor 2^k x \rfloor + 1$, and therefore $P[S_{nk} \le x] = (\lfloor 2^k x \rfloor + 1)/2^k$. Since $S_{nk}(\omega) \uparrow U_n(\omega)$ as $k \uparrow \infty$, it follows that $[S_{nk} \le x] \downarrow [U_n \le x]$ as $k \uparrow \infty$, and so $P[U_n \le x] = \lim_k P[S_{nk} \le x] = \lim_k (\lfloor 2^k x \rfloor + 1)/2^k = x$ for $0 \le x < 1$. Thus U_n is uniformly distributed over the unit interval.

The construction thus far establishes the existence of an independent sequence of random variables U_n each uniformly distributed over $[0,1]$. Let F_n be the distribution function corresponding to μ_n, and put $\varphi_n(u) = \inf[x: u \leq F_n(x)]$ for $0 < u < 1$. This is the inverse used in Section 14—see (14.5). Set $\varphi_n(u) = 0$, say, for u outside $(0,1)$, and put $X_n(\omega) = \varphi_n(U_n(\omega))$. Since $\varphi_n(u) \leq x$ if and only if $u \leq F_n(x)$—see the argument following (14.5)—$P[X_n \leq x] = P[U_n \leq F_n(x)] = F_n(x)$. Thus X_n has distribution function F_n. And by Theorem 20.2, $X_1, X_2, \ldots$ are independent.   ∎

This theorem of course includes Theorem 5.3 as a special case, and its proof does not depend on the earlier result. Theorem 20.4 is a special case of Kolmogorov's existence theorem in Section 36.

Convolution

Let X and Y be independent random variables with distributions μ and ν. Apply (20.27) and (20.29) to the planar set $B = [(x,y): x + y \in H]$ with $H \in \mathcal{R}^1$:

$$(20.34) \qquad P[X + Y \in H] = \int_{-\infty}^{\infty} \nu(H - x)\mu(dx)$$

$$= \int_{-\infty}^{\infty} P[Y \in H - x]\mu(dx).$$

The *convolution* of μ and ν is the measure $\mu * \nu$ defined by

$$(20.35) \qquad (\mu * \nu)(H) = \int_{-\infty}^{\infty} \nu(H - x)\mu(dx), \qquad H \in \mathcal{R}^1.$$

If X and Y are independent and have distributions μ and ν, (20.34) shows that $X + Y$ has distribution $\mu * \nu$. Since addition of random variables is commutative and associative, the same is true of convolution: $\mu * \nu = \nu * \mu$ and $\mu * (\nu * \eta) = (\mu * \nu) * \eta$.

If F and G are the distribution functions corresponding to μ and ν, the distribution function corresponding to $\mu * \nu$ is denoted $F * G$. Taking $H = (-\infty, y]$ in (20.35) shows that

$$(20.36) \qquad (F * G)(y) = \int_{-\infty}^{\infty} G(y - x)\, dF(x).$$

(See (17.22) for the notation $dF(x)$.) If G has density g, then $G(y - x) = \int_{-\infty}^{y-x} g(s)\, ds = \int_{-\infty}^{y} g(t - x)\, dt$, and so the right side of (20.36) is $\int_{-\infty}^{y} [\int_{-\infty}^{\infty} g(t - x)\, dF(x)]\, dt$ by Fubini's theorem. Thus $F * G$ has density $F * g$, where

$$(20.37) \qquad (F * g)(y) = \int_{-\infty}^{\infty} g(y - x)\, dF(x);$$

this holds if G has density g. If, in addition, F has density f, (20.37) is denoted $f * g$ and reduces by (16.12) to

$$(20.38) \qquad (f * g)(y) = \int_{-\infty}^{\infty} g(y - x) f(x) \, dx.$$

This defines convolution for densities, and $\mu * \nu$ has density $f * g$ if μ and ν have densities f and g. The formula (20.38) can be used for many explicit calculations.

Example 20.5. Let $X_1, \ldots, X_k$ be independent random variables, each with the exponential density (20.10). Define g_k by

$$(20.39) \qquad g_k(x) = \alpha \frac{(\alpha x)^{k-1}}{(k-1)!} e^{-\alpha x}, \qquad x \geq 0, \quad k = 1, 2, \ldots;$$

put $g_k(x) = 0$ for $x \leq 0$. Now

$$(g_{k-1} * g_1)(y) = \int_0^y g_{k-1}(y - x) g_1(x) \, dx,$$

which reduces to $g_k(y)$. Thus $g_k = g_{k-1} * g_1$, and since g_1 coincides with (20.10), it follows by induction that the sum $X_1 + \cdots + X_k$ has density g_k. The corresponding distribution function is

$$(20.40) \quad G_k(x) = 1 - e^{-\alpha x} \sum_{i=0}^{k-1} \frac{(\alpha x)^i}{i!} = \sum_{i=k}^{\infty} e^{-\alpha x} \frac{(\alpha x)^i}{i!}, \qquad x \geq 0,$$

as follows by differentiation. ∎

Example 20.6. Suppose that X has the normal density (20.12) with $m = 0$ and that Y has the same density with τ in place of σ. If X and Y are independent, then $X + Y$ has density

$$\frac{1}{2 \pi \sigma \tau} \int_{-\infty}^{\infty} \exp \left[- \frac{(y - x)^2}{2 \sigma^2} - \frac{x^2}{2 \tau^2} \right] dx.$$

A change of variable $u = x(\sigma^2 + \tau^2)^{1/2} / \sigma \tau$ reduces this to

$$\frac{1}{\sqrt{2 \pi (\sigma^2 + \tau^2)}} \frac{1}{\sqrt{2 \pi}} \int_{-\infty}^{\infty} \exp \left[- \frac{1}{2} \left(u - y \frac{\tau / \sigma}{\sqrt{\sigma^2 + \tau^2}} \right)^2 - \frac{y^2}{2(\sigma^2 + \tau^2)} \right] du$$

$$= \frac{1}{\sqrt{2 \pi (\sigma^2 + \tau^2)}} e^{-y^2 / 2(\sigma^2 + \tau^2)}.$$

Thus $X + Y$ has the normal density with $m = 0$ and with $\sigma^2 + \tau^2$ in place of σ^2. ∎

If μ and ν are arbitrary finite measures on the line, their convolution is defined by (20.35) even if they are not probability measures.

Convergence in Probability

Random variables X_n converge in probability to X, written $X_n \to_p X$, if

$$(20.41) \qquad \lim_n P[|X_n - X| \geq \epsilon] = 0$$

for each positive ϵ.[†] This is the same as (5.7), and the proof of Theorem 5.2 carries over without change (see also Example 5.4)

Theorem 20.5.   (i) *If $X_n \to X$ with probability* 1, *then $X_n \to_p X$.*

(ii) *A necessary and sufficient condition for $X_n \to_p X$ is that each subsequence $\{X_{n_k}\}$ contain a further subsequence $\{X_{n_{k(i)}}\}$ such that $X_{n_{k(i)}} \to X$ with probability* 1 *as $i \to \infty$*

PROOF.   Only part (ii) needs proof. If $X_n \to_p X$, then given $\{n_k\}$, choose a subsequence $\{n_{k(i)}\}$ so that $k \geq k(i)$ implies that $P[|X_{n_k} - X| \geq i^{-1}] < 2^{-i}$. By the first Borel–Cantelli lemma there is probability 1 that $|X_{n_{k(i)}} - X| < i^{-1}$ for all but finitely many i. Therefore, $\lim_i X_{n_{k(i)}} = X$ with probability 1.

If X_n does not converge to X in probability, there is some positive ϵ for which $P[|X_{n_k} - X| \geq \epsilon] > \epsilon$ holds along some sequence $\{n_k\}$. No subsequence of $\{X_{n_k}\}$ can converge to X in probability, and hence none can converge to X with probability 1. ∎

It follows from (ii) that if $X_n \to_p X$ and $X_n \to_p Y$, then $X = Y$ with probability 1. It follows further that if f is continuous and $X_n \to_p X$, then $f(X_n) \to_p f(X)$.

In nonprobabilistic contexts, convergence in probability becomes *convergence in measure*: If f_n and f are real measurable functions on a measure space $(\Omega, \mathcal{F}, \mu)$, and if $\mu[\omega: |f(\omega) - f_n(\omega)| \geq \epsilon] \to 0$ for each $\epsilon > 0$, then f_n converges in measure to f.

The Glivenko–Cantelli Theorem*

The *empirical distribution function* for random variables $X_1, \ldots, X_n$ is the distribution function $F_n(x, \omega)$ with a jump of n^{-1} at each $X_k(\omega)$:

$$(20.42) \qquad F_n(x, \omega) = \frac{1}{n} \sum_{k=1}^{n} I_{(-\infty, x]}(X_k(\omega)).$$

[†]This is often expressed $\mathrm{p\,lim}_n X_n = X$
*This topic may be omitted

If the X_k have a common unknown distribution function $F(x)$, then $F_n(x, \omega)$ is its natural estimate. The estimate has the right limiting behavior, according to the *Glivenko–Cantelli theorem*:

Theorem 20.6. *Suppose that $X_1, X_2, \ldots$ are independent and have a common distribution function F; put $D_n(\omega) = \sup_x |F_n(x, \omega) - F(x)|$. Then $D_n \to 0$ with probability* 1.

For each x, $F_n(x, \omega)$ as a function of ω is a random variable. By right continuity, the supremum above is unchanged if x is restricted to the rationals, and therefore D_n is a random variable.

The summands in (20.42) are independent, identically distributed simple random variables, and so by the strong law of large numbers (Theorem 6.1), for each x there is a set A_x of probability 0 such that

$$(20.43) \qquad \lim_n F_n(x, \omega) = F(x)$$

for $\omega \notin A_x$. But Theorem 20.6 says more, namely that (20.43) holds for ω outside some set A of probability 0, where A does not depend on x—as there are uncountably many of the sets A_x, it is conceivable a priori that their union might necessarily have positive measure. Further, the convergence in (20.43) is uniform in x. Of course, the theorem implies that with probability 1 there is weak convergence $F_n(x, \omega) \Rightarrow F(x)$ in the sense of Section 14.

PROOF OF THE THEOREM. As already observed, the set A_x where (20.43) fails has probability 0. Another application of the strong law of large numbers, with $I_{(-\infty, x)}$ in place of $I_{(-\infty, x]}$ in (20.42), shows that (see (20.5)) $\lim_n F_n(x - , \omega) = F(x -)$ except on a set B_x of probability 0. Let $\varphi(u) = \inf[x: u \le F(x)]$ for $0 < u < 1$ (see (14.5)), and put $x_{m,k} = \varphi(k/m)$, $m \ge 1$, $1 \le k \le m$. It is not hard to see that $F(\varphi(u) -) \le u \le F(\varphi(u))$; hence $F(x_{m,k} -) - F(x_{m,k-1}) \le m^{-1}$, $F(x_{m,1} -) \le m^{-1}$, and $F(x_{m,m}) \ge 1 - m^{-1}$. Let $D_{m,n}(\omega)$ be the maximum of the quantities $|F_n(x_{m,k}, \omega) - F(x_{m,k})|$ and $|F_n(x_{m,k} - , \omega) - F(x_{m,k} -)|$ for $k = 1, \ldots, m$.

If $x_{m,k-1} \le x < x_{m,k}$, then $F_n(x, \omega) \le F_n(x_{m,k} - , \omega) \le F(x_{m,k} -) + D_{m,n}(\omega) \le F(x) + m^{-1} + D_{m,n}(\omega)$ and $F_n(x, \omega) \ge F_n(x_{m,k-1}, \omega) \ge F(x_{m,k-1}) - D_{m,n}(\omega) \ge F(x) - m^{-1} - D_{m,n}(\omega)$. Together with similar arguments for the cases $x < x_{m,1}$ and $x \ge x_{m,m}$, this shows that

$$(20.44) \qquad D_n(\omega) \le D_{m,n}(\omega) + m^{-1}.$$

If ω lies outside the union A of all the $A_{x_{mk}}$ and $B_{x_{mk}}$, then $\lim_n D_{m,n}(\omega) = 0$ and hence $\lim_n D_n(\omega) = 0$ by (20.44). But A has probability 0. ■

PROBLEMS

20.1. 2.11 ↑ A necessary and sufficient condition for a σ-field $\mathscr{G}$ to be countably generated is that $\mathscr{G} = \sigma(X)$ for some random variable X. *Hint:* If $\mathscr{G} = \sigma(A_1, A_2, \ldots)$, consider $X = \sum_{k=1}^{\infty} f(I_{A_k})/10^k$, where $f(x)$ is 4 for $x = 0$ and 5 for $x \neq 0$.

20.2. If X is a positive random variable with density f, then X^{-1} has density $f(1/x)/x^2$. Prove this by (20.16) and by a direct argument.

20.3. Suppose that a two-dimensional distribution function F has a continuous density f. Show that $f(x, y) = \partial^2 F(x, y)/\partial x \, \partial y$.

20.4. The construction in Theorem 20.4 requires only Lebesgue measure on the unit interval. Use the theorem to prove the existence of Lebesgue measure on R^k. First construct λ_k restricted to $(-n, n] \times \cdots \times (-n, n]$, and then pass to the limit $(n \to \infty)$. The idea is to argue from first principles, and not to use previous constructions, such as those in Theorems 12.5 and 18.2.

20.5. Suppose that A, B, and C are positive, independent random variables with distribution function F. Show that the quadratic $Az^2 + Bz + C$ has real zeros with probability $\int_0^{\infty} \int_0^{\infty} F(x^2/4y) \, dF(x) \, dF(y)$.

20.6. Show that $X_1, X_2, \ldots$ are independent if $\sigma(X_1, \ldots, X_{n-1})$ and $\sigma(X_n)$ are independent for each n.

20.7. Let $X_0, X_1, \ldots$ be a persistent, irreducible Markov chain, and for a fixed state j let $T_1, T_2, \ldots$ be the times of the successive passages through j. Let $Z_1 = T_1$ and $Z_n = T_n - T_{n-1}$, $n \geq 2$. Show that $Z_1, Z_2, \ldots$ are independent and that $P[Z_n = k] = f_{jj}^{(k)}$ for $n \geq 2$.

20.8. *Ranks and records.* Let $X_1, X_2, \ldots$ be independent random variables with a common continuous distribution function. Let B be the ω-set where $X_m(\omega) = X_n(\omega)$ for some pair m, n of distinct integers, and show that $P(B) = 0$. Remove B from the space Ω on which the X_n are defined. This leaves the joint distributions of the X_n unchanged and makes ties impossible.

Let $T^{(n)}(\omega) = (T_1^{(n)}(\omega), \ldots, T_n^{(n)}(\omega))$ be that permutation $(t_1, \ldots, t_n)$ of $(1, \ldots, n)$ for which $X_{t_1}(\omega) < X_{t_2}(\omega) < \cdots < X_{t_n}(\omega)$. Let Y_n be the rank of X_n among $X_1, \ldots, X_n$: $Y_n = r$ if and only if $X_i < X_n$ for exactly $r - 1$ values of i preceding n.

(a) Show that $T^{(n)}$ is uniformly distributed over the $n!$ permutations.

(b) Show that $P[Y_n = r] = 1/n$, $1 \leq r \leq n$.

(c) Show that Y_k is measurable $\sigma(T^{(n)})$ for $k \leq n$.

(d) Show that $Y_1, Y_2, \ldots$ are independent.

20.9. ↑ *Record values.* Let A_n be the event that a *record* occurs at time n: $\max_{k < n} X_k < X_n$.

(a) Show that $A_1, A_2, \ldots$ are independent and $P(A_n) = 1/n$.

(b) Show that no record stands forever.

(c) Let N_n be the time of the first record after time n. Show that $P[N_n = n + k] = n(n + k - 1)^{-1}(n + k)^{-1}$.

20.10. Use Fubini's theorem to prove that convolution of finite measures is commutative and associative.

20.11. Suppose that X and Y are independent and have densities. Use (20.20) to find the joint density for $(X + Y, X)$ and then use (20.19) to find the density for $X + Y$. Check with (20.38).

20.12. If $F(x - \epsilon) < F(x + \epsilon)$ for all positive ϵ, then x is a *point of increase* of F (see Problem 12.9). If $F(x -) < F(x)$, then x is an *atom* of F.

(a) Show that, if x and y are points of increase of F and G, then $x + y$ is a point of increase of $F * G$.

(b) Show that, if x and y are atoms of F and G, then $x + y$ is an atom of $F * G$.

20.13. Suppose that μ and ν consist of masses α_n and β_n at n, $n = 0, 1, 2, \ldots$. Show that $\mu * \nu$ consists of a mass of $\sum_{k=0}^{n} \alpha_k \beta_{n-k}$ at n, $n = 0, 1, 2, \ldots$. Show that two Poisson distributions (the parameters may differ) convolve to a Poisson distribution.

20.14. The *Cauchy* distribution has density

$$(20.45) \qquad c_u(x) = \frac{1}{\pi} \frac{u}{u^2 + x^2}, \qquad -\infty < x < \infty,$$

for $u > 0$. (By (17.9), the density integrates to 1.)

(a) Show that $c_u * c_t = c_{u+t}$. *Hint:* Expand the convolution integrand in partial fractions.

(b) Show that, if $X_1, \ldots, X_n$ are independent and have density c_u, then $(X_1 + \cdots + X_n)/n$ has density c_u as well.

20.15. ↑ (a) Show that, if X and Y are independent and have the standard normal density, then X/Y has the Cauchy density with $u = 1$.

(b) Show that, if X has the uniform distribution over $(-\pi/2, \pi/2)$, then $\tan X$ has the Cauchy distribution with $u = 1$.

20.16. 18.18↑ Let $X_1, \ldots, X_n$ be independent, each having the standard normal distribution Show that

$$\chi_n^2 = X_1^2 + \cdots + X_n^2$$

has density

$$(20.46) \qquad \frac{1}{2^{n/2} \Gamma(n/2)} x^{(n/2)-1} e^{-x/2}$$

over $(0, \infty)$. This is called the *chi-squared distribution with n degrees of freedom*.

20.17. ↑   The *gamma distribution* has density

(20.47) $$f(x;\alpha,u) = \frac{\alpha^u}{\Gamma(u)} x^{u-1} e^{-\alpha x}$$

over $(0,\infty)$ for positive parameters α and u. Check that (20.47) integrates to 1 Show that

(20.48) $$f(\ ,\alpha,u) * f(\ ;\alpha,v) = f(\cdot;\alpha,u+v).$$

Note that (20.46) is $f(x;\frac{1}{2},n/2)$, and from (20.48) deduce again that (20.46) is the density of χ_n^2. Note that the exponential density (20 10) is $f(x;\alpha,1)$, and from (20 48) deduce (20 39) once again.

20.18. ↑   Let $N, X_1, X_2, \ldots$ be independent, where $P[N=n] = q^{n-1}p$, $n \geq 1$, and each X_k has the exponential density $f(x;\alpha,1)$. Show that $X_1 + \cdots + X_N$ has density $f(x;\alpha p,1)$.

20.19. Let $A_{nm}(\epsilon) = [|Z_k - Z| < \epsilon, n \leq k \leq m]$. Show that $Z_n \to Z$ with probability 1 if and only if $\lim_n \lim_m P(A_{nm}(\epsilon)) = 1$ for all positive ϵ, whereas $Z_n \to_p Z$ if and only if $\lim_n P(A_{nn}(\epsilon)) = 1$ for all positive ϵ.

20.20. (a) Suppose that $f: R^2 \to R^1$ is continuous. Show that $X_n \to_p X$ and $Y_n \to_p Y$ imply $f(X_n, Y_n) \to_p f(X,Y)$.

(b) Show that addition and multiplication preserve convergence in probability.

20.21. Suppose that the sequence $\{X_n\}$ is *fundamental in probability* in the sense that for ϵ positive there exists an N_ϵ such that $P[|X_m - X_n| > \epsilon] < \epsilon$ for $m,n > N_\epsilon$.

(a) Prove there is a subsequence $\{X_{n_k}\}$ and a random variable X such that $\lim_k X_{n_k} = X$ with probability 1. *Hint:* Choose increasing n_k such that $P[|X_m - X_n| > 2^{-k}] < 2^{-k}$ for $m,n \geq n_k$. Analyze $P[|X_{n_{k+1}} - X_{n_k}| > 2^{-k}]$.

(b) Show that $X_n \to_p X$.

20.22. (a) Suppose that $X_1 \leq X_2 \leq \cdots$ and that $X_n \to_p X$. Show that $X_n \to X$ with probability 1.

(b) Show by example that in an infinite measure space functions can converge almost everywhere without converging in measure.

20.23. If $X_n \to 0$ with probability 1, then $n^{-1} \sum_{k=1}^n X_k \to 0$ with probability 1 by the standard theorem on Cesàro means [A30]. Show by example that this is not so if convergence with probability 1 is replaced by convergence in probability.

20.24. 2.19↑   (a) Show that in a discrete probability space convergence in probability is equivalent to convergence with probability 1.

(b) Show that discrete spaces are essentially the only ones where this equivalence holds: Suppose that P has a nonatomic part in the sense that there is a set A such that $P(A) > 0$ and $P(\ |A)$ is nonatomic. Construct random variables X_n such that $X_n \to_p 0$ but X_n does not converge to 0 with probability 1.

20.25. 20.21 20.24↑ Let $d(X,Y)$ be the infimum of those positive ϵ for which $P[|X - Y| \geq \epsilon] \leq \epsilon$.

(a) Show that $d(X,Y) = 0$ if and only if $X = Y$ with probability 1. Identify random variables that are equal with probability 1, and show that d is a metric on the resulting space.

(b) Show that $X_n \rightarrow_p X$ if and only if $d(X_n, X) \rightarrow 0$.

(c) Show that the space is complete.

(d) Show that in general there is no metric d_0 on this space such that $X_n \rightarrow X$ with probability 1 if and only if $d_0(X_n, X) \rightarrow 0$.

20.26. Construct in R^k a random variable X that is uniformly distributed over the surface of the unit sphere in the sense that $|X| = 1$ and UX has the same distribution as X for orthogonal transformations U. *Hint* Let Z be uniformly distributed in the unit ball in R^k, define $\psi(x) = x/|x|$ ($\psi(0) = (1, 0, \quad, 0)$, say), and take $X = \psi(Z)$.

20.27. ↑ Let Θ and Φ be the longitude and latitude of a random point on the surface of the unit sphere in R^3. Show that Θ and Φ are independent, Θ is uniformly distributed over $[0, 2\pi)$, and Φ is distributed over $[-\pi/2, +\pi/2]$ with density $\frac{1}{2} \cos \phi$.

SECTION 21. EXPECTED VALUES

Expected Value as Integral

The expected value of a random variable X on $(\Omega, \mathscr{F}, P)$ is the integral of X with respect to the measure P:

$$E[X] = \int X dP = \int_{\Omega} X(\omega) P(d\omega).$$

All the definitions, conventions, and theorems of Chapter 3 apply. For nonnegative X, $E[X]$ is always defined (it may be infinite); for the general X, $E[X]$ is defined, or X has an expected value, if at least one of $E[X^+]$ and $E[X^-]$ is finite, in which case $E[X] = E[X^+] - E[X^-]$; and X is integrable if and only if $E[|X|] < \infty$. The integral $\int_A X dP$ over a set A is defined, as before, as $E[I_A X]$. In the case of simple random variables, the definition reduces to that used in Sections 5 through 9.

Expected Values and Limits

The theorems on integration to the limit in Section 16 apply. A useful fact: If random variables X_n are dominated by an integrable random variable, or if they are uniformly integrable, then $E[X_n] \rightarrow E[X]$ follows if X_n converges to X *in probability*—convergence with probability 1 is not necessary. This follows easily from Theorem 20.5.

Expected Values and Distributions

Suppose that X has distribution μ. If g is a real function of a real variable, then by the change-of-variable formula (16.17),

$$(21.1) \qquad E[g(X)] = \int_{-\infty}^{\infty} g(x)\mu(dx).$$

(In applying (16.17), replace $T\colon \Omega \to \Omega'$ by $X\colon \Omega \to R^1$, μ by P, μT^{-1} by μ, and f by g.) This formula holds in the sense explained in Theorem 16.13: It holds in the nonnegative case, so that

$$(21.2) \qquad E[|g(X)|] = \int_{-\infty}^{\infty} |g(x)|\mu(dx);$$

if one side is infinite, then so is the other. And if the two sides of (21.2) are finite, then (21.1) holds.

If μ is discrete and $\mu\{x_1, x_2, \ldots\} = 1$, then (21.1) becomes (use Theorem 16.9)

$$(21.3) \qquad E[g(X)] = \sum_r g(x_r)\mu\{x_r\}.$$

If X has density f, then (21.1) becomes (use Theorem 16.11)

$$(21.4) \qquad E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)\, dx.$$

If F is the distribution function of X and μ, (21.1) can be written $E[g(X)] = \int_{-\infty}^{\infty} g(x)\, dF(x)$ in the notation (17.22).

Moments

By (21.2), μ and F determine all the *absolute moments* of X:

$$(21.5) \quad E[|X|^k] = \int_{-\infty}^{\infty} |x|^k \mu(dx) = \int_{-\infty}^{\infty} |x|^k\, dF(x), \qquad k = 1, 2, \ldots .$$

Since $j \le k$ implies that $|x|^j \le 1 + |x|^k$, if X has a finite absolute moment of order k, then it has finite absolute moments of orders $1, 2, \ldots, k-1$ as well. For each k for which (2.15) is finite, X has kth *moment*

$$(21.6) \qquad E[X^k] = \int_{-\infty}^{\infty} x^k \mu(dx) = \int_{-\infty}^{\infty} x^k\, dF(x).$$

These quantities are also referred to as the moments of μ and of F. They can be computed by (21.3) and (21.4) in the appropriate circumstances.

Example 21.1. Consider the normal density (20.12) with $m = 0$ and $\sigma = 1$. For each k, $x^k e^{-x^2/2}$ goes to 0 exponentially as $x \to \pm\infty$, and so finite moments of all orders exist. Integration by parts shows that

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^k e^{-x^2/2} \, dx = \frac{k-1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^{k-2} e^{-x^2/2} \, dx, \qquad k = 2, 3, \ldots .$$

(Apply (18.16) to $g(x) = x^{k-2}$ and $f(x) = x e^{-x^2/2}$, and let $a \to -\infty$, $b \to \infty$.) Of course, $E[X] = 0$ by symmetry and $E[X^0] = 1$. It follows by induction that

$$(21.7) \qquad E\left[X^{2k}\right] = 1 \times 3 \times 5 \times \cdots \times (2k-1), \qquad k = 1, 2, \ldots,$$

and that the odd moments all vanish.   ∎

If the first two moments of X are finite and $E[X] = m$, then just as in Section 5, the *variance* is

$$(21.8) \qquad \mathrm{Var}[X] = E\left[(X-m)^2\right] = \int_{-\infty}^{\infty} (x-m)^2 \mu(dx)$$

$$= E\left[X^2\right] - m^2.$$

From Example 21.1 and a change of variable, it follows that a random variable with the normal density (20.12) has mean m and variance σ^2.

Consider for nonnegative X the relation

$$(21.9) \qquad E[X] = \int_0^{\infty} P[X > t] \, dt = \int_0^{\infty} P[X \geq t] \, dt.$$

Since $P[X = t]$ can be positive for at most countably many values of t, the two integrands differ only on a set of Lebesgue measure 0 and hence the integrals are the same. For X simple and nonnegative, (21.9) was proved in Section 5; see (5.29). For the general nonnegative X, let X_n be simple random variables for which $0 \leq X_n \uparrow X$ (see (20.1)). By the monotone convergence theorem $E[X_n] \uparrow E[X]$; moreover, $P[X_n > t] \uparrow P[X > t]$, and therefore $\int_0^{\infty} P[X_n > t] \, dt \uparrow \int_0^{\infty} P[X > t] \, dt$, again by the monotone convergence theorem. Since (21.9) holds for each X_n, a passage to the limit establishes (21.9) for X itself. Note that both sides of (21.9) may be infinite. If the integral on the right is finite, then X is integrable.

Replacing X by $XI_{[X > \alpha]}$ leads from (21.9) to

$$(21.10) \qquad \int_{[X > \alpha]} X \, dP = \alpha P[X > \alpha] + \int_{\alpha}^{\infty} P[X > t] \, dt, \qquad \alpha \geq 0.$$

As long as $\alpha \geq 0$, this holds even if X is not nonnegative.

Inequalities

Since the final term in (21.10) is nonnegative, $\alpha P[X \geq \alpha] \leq \int_{[X \geq \alpha]} X dP \leq E[X]$. Thus

$$(21.11) \qquad P[X \geq \alpha] \leq \frac{1}{\alpha} \int_{[X \geq \alpha]} X dP \leq \frac{1}{\alpha} E[X], \qquad \alpha > 0,$$

for nonnegative X. As in Section 5, there follow the inequalities

$$(21.12) \qquad P[|X| \geq \alpha] \leq \frac{1}{\alpha^k} \int_{[|X| \geq \alpha]} |X|^k \, dP \leq \frac{1}{\alpha^k} E[|X|^k].$$

It is the inequality between the two extreme terms here that usually goes under the name of Markov; but the left-hand inequality is often useful, too. As a special case there is Chebyshev's inequality,

$$(21.13) \qquad\qquad P[|X - m| \geq \alpha] \leq \frac{1}{\alpha^2} \operatorname{Var}[X]$$

$(m = E[X])$.

Jensen's inequality

$$(21.14) \qquad\qquad \varphi(E[X]) \leq E[\varphi(X)]$$

holds if φ is convex on an interval containing the range of X and if X and $\varphi(X)$ both have expected values. To prove it, let $l(x) = ax + b$ be a supporting line through $(E[X], \varphi(E[X]))$—a line lying entirely under the graph of φ [A33]. Then $aX(\omega) + b \leq \varphi(X(\omega))$, so that $aE[X] + b \leq E[\varphi(X)]$. But the left side of this inequality is $\varphi(E[X])$.

Hölder's inequality is

$$(21.15) \qquad E[|XY|] \leq E^{1/p}[|X|^p] E^{1/q}[|Y|^q], \qquad \frac{1}{p} + \frac{1}{q} = 1.$$

For discrete random variables, this was proved in Section 5; see (5.35). For the general case, choose simple random variables X_n and Y_n satisfying $0 \leq |X_n| \uparrow |X|$ and $0 \leq |Y_n| \uparrow |Y|$; see (20.2). Then (5.35) and the monotone convergence theorem give (21.15). Notice that (21.15) implies that if $|X|^p$ and $|Y|^q$ are integrable, then so is XY. Schwarz's inequality is the case $p = q = 2$:

$$(21.16) \qquad\qquad E[|XY|] \leq E^{1/2}[X^2] E^{1/2}[Y^2].$$

If X and Y have second moments, then XY must have a first moment.

The same reasoning shows that Lyapounov's inequality (5.37) carries over from the simple to the general case.

Joint Integrals

The relation (21.1) extends to random vectors. Suppose that $(X_1, \ldots, X_k)$ has distribution μ in k-space and $g: R^k \to R^1$. By Theorem 16.13,

$$(21.17) \qquad E[g(X_1, \ldots, X_k)] = \int_{R^k} g(x)\mu(dx),$$

with the usual provisos about infinite values. For example, $E[X_i X_j] = \int_{R^k} x_i x_j \mu(dx)$. If $E[X_i] = m_i$, the *covariance* of X_i and X_j is

$$\text{Cov}[X_i, X_j] = E[(X_i - m_i)(X_j - m_j)] = \int_{R^k} (x_i - m_i)(x_j - m_j)\mu(dx).$$

Random variables are *uncorrelated* if they have covariance 0.

Independence and Expected Value

Suppose that X and Y are independent. If they are also simple, then $E[XY] = E[X]E[Y]$, as proved in Section 5—see (5.25). Define X_n by (20.2) and similarly define $Y_n = \psi_n(Y^+) - \psi_n(Y^-)$. Then X_n and Y_n are independent and simple, so that $E[|X_n Y_n|] = E[|X_n|]E[|Y_n|]$, and $0 \le |X_n| \uparrow |X|, 0 \le |Y_n| \uparrow |Y|$. If X and Y are integrable, then $E[|X_n Y_n|] = E[|X_n|]E[|Y_n|] \uparrow E[|X|] \cdot E[|Y|]$, and it follows by the monotone convergence theorem that $E[|XY|] < \infty$; since $X_n Y_n \to XY$ and $|X_n Y_n| \le |XY|$, it follows further by the dominated convergence theorem that $E[XY] = \lim_n E[X_n Y_n] = \lim_n E[X_n]E[Y_n] = E[X]E[Y]$. Therefore, XY is integrable if X and Y are (which is by no means true for dependent random variables) and $E[XY] = E[X]E[Y]$.

This argument obviously extends inductively: If $X_1, \ldots, X_k$ are independent and integrable, then the product $X_1 \cdots X_k$ is also integrable and

$$(21.18) \qquad E[X_1 \cdots X_k] = E[X_1] \cdots E[X_k].$$

Suppose that $\mathscr{G}_1$ and $\mathscr{G}_2$ are independent σ-fields, A lies in $\mathscr{G}_1$, X_1 is measurable $\mathscr{G}_1$, and X_2 is measurable $\mathscr{G}_2$. Then $I_A X_1$ and X_2 are independent, so that (21.18) gives

$$(21.19) \qquad \int_A X_1 X_2 \, dP = \int_A X_1 \, dP \cdot E[X_2]$$

if the random variables are integrable. In particular,

$$(21.20) \qquad \int_A X_2 \, dP = P(A)E[X_2].$$

From (21.18) it follows just as for simple random variables (see (5.28)) that variances add for sums of independent random variables. It is even enough that the random variables be independent in pairs.

Moment Generating Functions

The *moment generating function* is defined as

$$(21.21) \qquad M(s) = E\big[e^{sX}\big] = \int_{-\infty}^{\infty} e^{sx}\mu(dx) = \int_{-\infty}^{\infty} e^{sx}\,dF(x)$$

for all s for which this is finite (note that the integrand is nonnegative). Section 9 shows in the case of simple random variables the power of moment generating function methods. This function is also called the *Laplace transform* of μ, especially in nonprobabilistic contexts.

Now $\int_0^\infty e^{sx}\mu(dx)$ is finite for $s \le 0$, and if it is finite for a positive s, then it is finite for all smaller s. Together with the corresponding result for the left half-line, this shows that $M(s)$ is defined on some interval containing 0. If X is nonnegative, this interval contains $(-\infty, 0]$ and perhaps part of $(0, \infty)$; if X is nonpositive, it contains $[0, \infty)$ and perhaps part of $(-\infty, 0)$. It is possible that the interval consists of 0 alone; this happens, for example, if μ is concentrated on the integers and $\mu\{n\} = \mu\{-n\} = C/n^2$ for $n = 1, 2, \ldots$.

Suppose that $M(s)$ is defined throughout an interval $(-s_0, s_0)$, where $s_0 > 0$. Since $e^{|sx|} \le e^{sx} + e^{-sx}$ and the latter function is integrable μ for $|s| < s_0$, so is $\sum_{k=0}^{\infty} |sx|^k/k! = e^{|sx|}$. By the corollary to Theorem 16.7, μ has finite moments of all orders and

$$(21.22) \quad M(s) = \sum_{k=0}^{\infty} \frac{s^k}{k!} E\big[X^k\big] = \sum_{k=0}^{\infty} \frac{s^k}{k!} \int_{-\infty}^{\infty} x^k \mu(dx), \qquad |s| < s_0.$$

Thus $M(s)$ has a Taylor expansion about 0 with positive radius of convergence if it is defined in some $(-s_0, s_0)$, $s_0 > 0$. If $M(s)$ can somehow be calculated and expanded in a series $\sum_k a_k s^k$, and if the coefficients a_k can be identified, then, since a_k must coincide with $E[X^k]/k!$, the moments of X can be computed: $E[X^k] = a_k k!$ It also follows from the theory of Taylor expansions [A29] that $a_k k!$ is the kth derivative $M^{(k)}(s)$ evaluated at $s = 0$:

$$(21.23) \qquad M^{(k)}(0) = E\big[X^k\big] = \int_{-\infty}^{\infty} x^k \mu(dx).$$

This holds if $M(s)$ exists in some neighborhood of 0.

Suppose now that M is defined in some neighborhood of s. If ν has density $e^{sx}/M(s)$ with respect to μ (see (16.11)), then ν has moment generating function $N(u) = M(s + u)/M(s)$ for u in some neighborhood of 0.

But then by (21.23), $N^{(k)}(0) = \int_{-\infty}^{\infty} x^k \nu(dx) = \int_{-\infty}^{\infty} x^k e^{sx} \mu(dx)/M(s)$, and since $N^{(k)}(0) = M^{(k)}(s)/M(s)$,

$$(21.24) \qquad M^{(k)}(s) = \int_{-\infty}^{\infty} x^k e^{sx} \mu(dx).$$

This holds as long as the moment generating function exists in some neighborhood of s. If $s = 0$, this gives (21.23) again. Taking $k = 2$ shows that $M(s)$ is convex in its interval of definition.

Example 21.2. For the standard normal density,

$$M(s) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{sx} e^{-x^2/2} dx = \frac{1}{\sqrt{2\pi}} e^{s^2/2} \int_{-\infty}^{\infty} e^{-(x-s)^2/2} dx,$$

and a change of variable gives

$$(21.25) \qquad M(s) = e^{s^2/2}.$$

The moment generating function in this case defined for all s. Since $e^{s^2/2}$ has the expansion

$$e^{s^2/2} = \sum_{k=0}^{\infty} \frac{1}{k!} \left(\frac{s^2}{2} \right)^k = \sum_{k=0}^{\infty} \frac{1 \times 3 \times \cdots \times (2k-1)}{(2k)!} s^{2k},$$

the moments can be read off from (21.22), which proves (21.7) once more. ∎

Example 21.3. In the exponential case (20.10), the moment generating function

$$(21.26) \qquad M(s) = \int_0^{\infty} e^{sx} \alpha e^{-\alpha x} dx = \frac{\alpha}{\alpha - s} = \sum_{k=0}^{\infty} \frac{s^k}{\alpha^k}$$

is defined for $s < \alpha$. By (21.22) the kth moment is $k!\alpha^{-k}$. The mean and variance are thus α^{-1} and α^{-2}. ∎

Example 21.4. For the Poisson distribution (20.7),

$$(21.27) \qquad M(s) = \sum_{r=0}^{\infty} e^{rs} e^{-\lambda} \frac{\lambda^r}{r!} = e^{\lambda(e^s - 1)}.$$

Since $M'(s) = \lambda e^s M(s)$ and $M''(s) = (\lambda^2 e^{2s} + \lambda e^s)M(s)$, the first two moments are $M'(0) = \lambda$ and $M''(0) = \lambda^2 + \lambda$; the mean and variance are both λ. ∎

Let $X_1, \ldots, X_k$ be independent random variables, and suppose that each X_i has a moment generating function $M_i(s) = E[e^{sX_i}]$ in $(-s_0, s_0)$. For $|s| < s_0$, each $\exp(sX_i)$ is integrable, and, since they are independent, their product $\exp(s\sum_{i=1}^{k} X_i)$ is also integrable (see (21.18)). The moment generating function of $X_1 + \cdots + X_k$ is therefore

$$(21\ 28) \qquad M(s) = M_1(s) \cdots M_k(s)$$

in $(-s_0, s_0)$. This relation for simple random variables was essential to the arguments in Section 9.

For simple random variables it was shown in Section 9 that the moment generating function determines the distribution. This will later be proved for general random variables; see Theorem 22.2 for the nonnegative case and Section 30 for the general case.

PROBLEMS

21.1. Prove

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-tx^2/2} \, dx = t^{-1/2},$$

differentiate k times with respect to t inside the integral (justify), and derive (21.7) again.

21.2. Show that, if X has the standard normal distribution, then $E[|X|^{2n+1}] = 2^n n! \sqrt{2/\pi}$.

21.3. 20.9↑   *Records*. Consider the sequence of records in the sense of Problem 20.9. Show that the expected waiting time to the next record is infinite.

21.4. 20.14↑   Show that the Cauchy distribution has no mean.

21.5. Prove the first Borel–Cantelli lemma by applying Theorem 16.6 to indicator random variables. Why is Theorem 16.6 not enough for the second Borel–Cantelli lemma?

21.6. Prove (21.9) by Fubini's theorem.

21.7. Prove for integrable X that

$$E[X] = \int_0^{\infty} P[X > t] \, dt - \int_{-\infty}^{0} P[X < t] \, dt.$$

21.8. (a) Suppose that X and Y have first moments, and prove

$$E[Y] - E[X] = \int_{-\infty}^{\infty} (P[X < t \leq Y] - P[Y < t \leq X]) \, dt.$$

(b) Let (X, Y) be a nondegenerate random interval Show that its expected length is the integral with respect to t of the probability that it covers t.

21.9. Suppose that X and Y are random variables with distribution functions F and G.

(a) Show that if F and G have no common jumps, then $E[F(Y)] + E[G(X)] = 1$.

(b) If F is continuous, then $E[F(X)] = \frac{1}{2}$.

(c) Even if F and G have common jumps, if X and Y are taken to be independent, then $E[F(Y)] + E[G(X)] = 1 + P[X = Y]$.

(d) Even if F has jumps, $E[F(X)] = \frac{1}{2} + \frac{1}{2}\sum_x P^2[X = x]$

21.10. (a) Show that uncorrelated variables need not be independent.

(b) Show that $\operatorname{Var}[\sum_{j=1}^{n} X_j] = \sum_{i,j=1}^{n} \operatorname{Cov}[X_i, X_j] = \sum_{i=1}^{n} \operatorname{Var}[X_i] + 2\sum_{1 \leq i < j \leq n} \operatorname{Cov}[X_i, X_j]$. The cross terms drop out if the X_i are uncorrelated, and hence drop out if they are independent.

21.11. ↑ Let X, Y, and Z be independent random variables such that X and Y assume the values $0, 1, 2$ with probability $\frac{1}{3}$ each and Z assumes the values 0 and 1 with probabilities $\frac{1}{3}$ and $\frac{2}{3}$. Let $X' = X$ and $Y' = X + Z$ (mod 3).

(a) Show that X', Y', and $X' + Y'$ have the same one-dimensional distributions as X, Y, and $X + Y$, respectively, even though (X', Y') and (X, Y) have different distributions.

(b) Show that X' and Y' are dependent but uncorrelated

(c) Show that, despite dependence, the moment generating function of $X' + Y'$ is the product of the moment generating functions of X' and Y'.

21.12. Suppose that X and Y are independent, nonnegative random variables and that $E[X] = \infty$ and $E[Y] = 0$. What is the value common to $E[XY]$ and $E[X]E[Y]$? Use the conventions (15.2) for both the product of the random variables and the product of their expected values. What if $E[X] = \infty$ and $0 < E[Y] < \infty$?

21.13. Suppose that X and Y are independent and that $f(x, y)$ is nonnegative. Put $g(x) = E[f(x, Y)]$ and show that $E[g(X)] = E[f(X, Y)]$. Show more generally that $\int_{X \in A} g(X) \, dP = \int_{X \in A} f(X, Y) \, dP$. Extend to f that may be negative.

21.14. ↑ The integrability of $X + Y$ does not imply that of X and Y separately. Show that it does if X and Y are independent.

21.15. 20.25 ↑ Write $d_1(X, Y) = E[|X - Y|/(1 + |X - Y|)]$. Show that this is a metric equivalent to the one in Problem 20.25.

21.16. For the density $C \exp(-|x|^{1/2})$, $-\infty < x < \infty$, show that moments of all orders exist but that the moment generating function exists only at $s = 0$.

21.17. 16.6 ↑ Show that a moment generating function $M(s)$ defined in $(-s_0, s_0)$, $s_0 > 0$, can be extended to a function analytic in the strip $[z: -s_0 < \text{Re } z < s_0]$. If $M(s)$ is defined in $[0, s_0)$, $s_0 > 0$, show that it can be extended to a function continuous in $[z: 0 \le \text{Re } z < s_0]$ and analytic in $[z: 0 < \text{Re } z < s_0]$.

21.18. Use (21.28) to find the generating function of (20.39).

21.19. For independent random variables having moment generating functions, show by (21.28) that the variances add.

21.20. 20.17 ↑ Show that the gamma density (20.47) has moment generating function $(1 - s/\alpha)^{-u}$ for $s < \alpha$. Show that the kth moment is $u(u + 1) \cdots (u + k - 1)/\alpha^k$. Show that the chi-squared distribution with n degrees of freedom has mean n and variance $2n$.

21.21. Let $X_1, X_2, \ldots$ be identically distributed random variables with finite second moment. Show that $nP[|X_1| \ge \epsilon\sqrt{n}] \to 0$ and $n^{-1/2} \max_{k \le n} |X_k| \to_p 0$.

SECTION 22. SUMS OF INDEPENDENT RANDOM VARIABLES

Let $X_1, X_2, \ldots$ be a sequence of independent random variables on some probability space. It is natural to ask whether the infinite series $\sum_{n=1}^{\infty} X_n$ converges with probability 1, or as in Section 6 whether $n^{-1}\sum_{k=1}^{n} X_k$ converges to some limit with probability 1. It is to questions of this sort that the present section is devoted.

Throughout the section, S_n will denote the partial sum $\sum_{k=1}^{n} X_k$ ($S_0 = 0$).

The Strong Law of Large Numbers

The central result is a general version of Theorem 6.1.

Theorem 22.1. *If $X_1, X_2, \ldots$ are independent and identically distributed and have finite mean, then $S_n/n \to E[X_1]$ with probability 1.*

Formerly this theorem stood at the end of a chain of results. The following argument, due to Etemadi, proceeds from first principles.

PROOF. If the theorem holds for nonnegative random variables, then $n^{-1}S_n = n^{-1}\sum_{k=1}^{n} X_k^+ - n^{-1}\sum_{k=1}^{n} X_k^- \to E[X_1^+] - E[X_1^-] = E[X_1]$ with probability 1. Assume then that $X_k \ge 0$.

Consider the truncated random variables $Y_k = X_k I_{[X_k \le k]}$ and their partial sums $S_n^* = \sum_{k=1}^{n} Y_k$. For $\alpha > 1$, temporarily fixed, let $u_n = \lfloor \alpha^n \rfloor$. The first step is to prove

$$(22.1) \qquad \sum_{n=1}^{\infty} P\left[\left| \frac{S_{u_n}^* - E[S_{u_n}^*]}{u_n} \right| > \epsilon \right] < \infty.$$

Since the X_n are independent and identically distributed,

$$\mathrm{Var}[S_n^*] = \sum_{k=1}^n \mathrm{Var}[Y_k] \le \sum_{k=1}^n E[Y_k^2]$$

$$= \sum_{k=1}^n E[X_1^2 I_{[X_1 \le k]}] \le nE[X_1^2 I_{[X_1 \le n]}].$$

It follows by Chebyshev's inequality that the sum in (22.1) is at most

$$\sum_{n=1}^\infty \frac{\mathrm{Var}[S_{u_n}^*]}{\epsilon^2 u_n^2} \le \frac{1}{\epsilon^2} E\left[X_1^2 \sum_{n=1}^\infty \frac{1}{u_n} I_{[X_1 \le u_n]} \right].$$

Let $K = 2\alpha/(\alpha - 1)$, and suppose $x > 0$. If N is the smallest n such that $u_n \ge x$, then $\alpha^N \ge x$, and since $y \le 2\lfloor y \rfloor$ for $y \ge 1$,

$$\sum_{u_n \ge x} u_n^{-1} \le 2 \sum_{n \ge N} \alpha^{-n} = K\alpha^{-N} \le Kx^{-1}.$$

Therefore, $\sum_{n=1}^\infty u_n^{-1} I_{[X_1 \le u_n]} \le KX_1^{-1}$ for $X_1 > 0$, and the sum in (22.1) is at most $K\epsilon^{-2} E[X_1] < \infty$.

From (22.1) it follows by the first Borel–Cantelli lemma (take a union over positive, rational ϵ) that $(S_{u_n}^* - E[S_{u_n}^*])/u_n \to 0$ with probability 1. But by the consistency of Cesàro summation [A30], $n^{-1}E[S_n^*] = n^{-1}\sum_{k=1}^n E[Y_k]$ has the same limit as $E[Y_n]$, namely. $E[X_1]$. Therefore $S_{u_n}^*/u_n \to E[X_1]$ with probability 1. Since

$$\sum_{n=1}^\infty P[X_n \ne Y_n] = \sum_{n=1}^\infty P[X_1 > n] \le \int_0^\infty P[X_1 > t]\, dt = E[X_1] < \infty,$$

another application of the first Borel–Cantelli lemma shows that $(S_n^* - S_n)/n \to 0$ and hence

(22.2)
$$\frac{S_{u_n}}{u_n} \to E[X_1]$$

with probability 1.

If $u_n \le k \le u_{n+1}$, then since $X_i \ge 0$,

$$\frac{u_n}{u_{n+1}} \frac{S_{u_n}}{u_n} \le \frac{S_k}{k} \le \frac{u_{n+1}}{u_n} \frac{S_{u_{n+1}}}{u_{n+1}}.$$

But $u_{n+1}/u_n \to \alpha$, and so it follows by (22.2) that

$$\frac{1}{\alpha}E[X_1] \le \lim_k \inf \frac{S_k}{k} \le \lim_k \sup \frac{S_k}{k} \le \alpha E[X_1]$$

with probability 1. This is true for each $\alpha > 1$. Intersecting the corresponding sets over rational α exceeding 1 gives $\lim_k S_k/k = E[X_1]$ with probability 1. ∎

Although the hypothesis that the X_n all have the same distribution is used several times in this proof, independence is used only through the equation $\mathrm{Var}[S_n^*] = \sum_{k=1}^n \mathrm{Var}[Y_k]$, and for this it is enough that the X_n be independent in pairs. The proof given for Theorem 6.1 of course extends beyond the case of simple random variables, but it requires $E[X_1^4] < \infty$.

Corollary. *Suppose that $X_1, X_2, \ldots$ are independent and identically distributed and $E[X_1^-] < \infty$, $E[X_1^+] = \infty$ (so that $E[X_1] = \infty$). Then $n^{-1}\sum_{k=1}^n X_k \to \infty$ with probability 1.*

PROOF. By the theorem, $n^{-1}\sum_{k=1}^n X_k^- \to E[X_1^-]$ with probability 1, and so it suffices to prove the corollary for the case $X_1 = X_1^+ \ge 0$. If

$$X_n^{(u)} = \begin{cases} X_n & \text{if } 0 \le X_n \le u, \\ 0 & \text{if } X_n > u, \end{cases}$$

then $n^{-1}\sum_{k=1}^n X_k \ge n^{-1}\sum_{k=1}^n X_k^{(u)} \to E[X_1^{(u)}]$ by the theorem. Let $u \to \infty$. ∎

The Weak Law and Moment Generating Functions

The weak law of large numbers (Section 6) carries over without change to the case of general random variables with second moments—only Chebyshev's inequality is required. The idea can be used to prove in a very simple way that a distribution concentrated on $[0, \infty)$ is uniquely determined by its moment generating function or Laplace transform.

For each λ, let Y_λ be a random variable (on some probability space) having the Poisson distribution with parameter λ. Since Y_λ has mean and variance λ (Example 21.4), Chebyshev's inequality gives

$$P\left[\left|\frac{Y_\lambda - \lambda}{\lambda}\right| \ge \epsilon\right] \le \frac{\lambda}{\lambda^2 \epsilon^2} \to 0, \qquad \lambda \to \infty.$$

Let G_λ be the distribution function of Y_λ/λ, so that

$$G_\lambda(t) = \sum_{k=0}^{\lfloor \lambda t \rfloor} e^{-\lambda} \frac{\lambda^k}{k!}.$$

The result above can be restated as

$$(22.3) \qquad \lim_{\lambda \to \infty} G_\lambda(t) = \begin{cases} 1 & \text{if } t > 1, \\ 0 & \text{if } t < 1. \end{cases}$$

In the notation of Section 14, $G_\lambda(x) \Rightarrow \Delta(x - 1)$ as $\lambda \to \infty$.

Now consider a probability distribution μ concentrated on $[0, \infty)$. Let F be the corresponding distribution function. Define

$$(22.4) \qquad M(s) = \int_0^\infty e^{-sx} \mu(dx), \qquad s \ge 0;$$

here 0 is included in the range of integration. This is the moment generating function (21.21), but the argument has been reflected through the origin. It is a *one-sided Laplace transform*, defined for all nonnegative s.

For positive s, (21.24) gives

$$(22.5) \qquad M^{(k)}(s) = (-1)^k \int_0^\infty y^k e^{-sy} \mu(dy).$$

Therefore, for positive x and s,

$$(22.6) \qquad \sum_{k=0}^{\lfloor sx \rfloor} \frac{(-1)^k}{k!} s^k M^{(k)}(s) = \int_0^\infty \sum_{k=0}^{\lfloor sx \rfloor} e^{-sy} \frac{(sy)^k}{k!} \mu(dy)$$

$$= \int_0^\infty G_{sy}\left(\frac{x}{y}\right) \mu(dy).$$

Fix $x > 0$. If[†] $0 \le y < x$, then $G_{sy}(x/y) \to 1$ as $s \to \infty$ by (22.3); if $y > x$, the limit is 0. If $\mu\{x\} = 0$, the integrand on the right in (22.6) thus converges as $s \to \infty$ to $I_{[0,x]}(y)$ except on a set of μ-measure 0. The bounded convergence theorem then gives

$$(22.7) \qquad \lim_{s \to \infty} \sum_{k=0}^{\lfloor sx \rfloor} \frac{(-1)^k}{k!} s^k M^{(k)}(s) = \mu[0, x] = F(x).$$

[†]If $y = 0$, the integrand in (22.5) is 1 for $k = 0$ and 0 for $k \ge 1$, hence for $y = 0$, the integrand in the middle term of (22 6) is 1.

Thus $M(s)$ determines the value of F at x if $x > 0$ and $\mu\{x\} = 0$, which covers all but countably many values of x in $[0, \infty)$. Since F is right-continuous, F itself and hence μ are determined through (22.7) by $M(s)$. In fact μ is by (22.7) determined by the values of $M(s)$ for s beyond an arbitrary s_0:

Theorem 22.2. *Let μ and ν be probability measures on $[0, \infty)$. If*

$$\int_0^\infty e^{-sx}\mu(dx) = \int_0^\infty e^{-sx}\nu(dx), \qquad s \geq s_0,$$

where $s_0 \geq 0$, then $\mu = \nu$.

Corollary. *Let f_1 and f_2 be real functions on $[0, \infty)$. If*

$$\int_0^\infty e^{-sx}f_1(x)\, dx = \int_0^\infty e^{-sx}f_2(x)\, dx, \qquad s \geq s_0,$$

where $s_0 \geq 0$, then $f_1 = f_2$ outside a set of Lebesgue measure 0.

The f_i need not be nonnegative, and they need not be integrable, but $e^{-sx}f_i(x)$ must be integrable over $[0, \infty)$ for $s \geq s_0$.

PROOF. For the nonnegative case, apply the theorem to the probability densities $g_i(x) = e^{-s_0 x}f_i(x)/m$, where $m = \int_0^\infty e^{-s_0 x}f_i(x)\, dx$, $i = 1, 2$. For the general case, prove that $f_1^+ + f_2^- = f_2^+ + f_1^-$ almost everywhere. ∎

Example 22.1. If $\mu_1 * \mu_2 = \mu_3$, then the corresponding transforms (22.4) satisfy $M_1(s)M_2(s) = M_3(s)$ for $s \geq 0$. If μ_i is the Poisson distribution with mean λ_i, then (see (21.27)) $M_i(s) = \exp[\lambda_i(e^{-s} - 1)]$. It follows by Theorem 22.2 that if two of the μ_i are Poisson, so is the third, and $\lambda_1 + \lambda_2 = \lambda_3$. ∎

Kolmogorov's Zero–One Law

Consider the set A of ω for which $n^{-1}\sum_{k=1}^n X_k(\omega) \to 0$ as $n \to \infty$. For each m, the values of $X_1(\omega), \ldots, X_{m-1}(\omega)$ are irrelevant to the question of whether or not ω lies in A, and so A ought to lie in the σ-field $\sigma(X_m, X_{m+1}, \ldots)$. In fact, $\lim_n n^{-1}\sum_{k=0}^{m-1} X_k(\omega) = 0$ for fixed m, and hence ω lies in A if and only if $\lim_n n^{-1}\sum_{k=m}^n X_k(\omega) = 0$. Therefore,

$$(22.8) \qquad A = \bigcap_\epsilon \bigcup_{N \geq m} \bigcap_{n \geq N} \left[\omega : \left| n^{-1} \sum_{k=m}^n X_k(\omega) \right| < \epsilon \right],$$

the first intersection extending over positive rational ϵ. The set on the inside

lies in $\sigma(X_m, X_{m+1}, \ldots)$, and hence so does A. Similarly, the ω-set where the series $\sum_n X_n(\omega)$ converges lies in each $\sigma(X_m, X_{m+1}, \ldots)$.

The intersection $\mathcal{T} = \bigcap_{n=1}^{\infty} \sigma(X_n, X_{n+1}, \ldots)$ is the *tail* σ-field associated with the sequence $X_1, X_2, \ldots$; its elements are *tail events*. In the case $X_n = I_{A_n}$, this is the σ-field (4.29) studied in Section 4. The following general form of *Kolmogorov's zero–one law* extends Theorem 4.5.

Theorem 22.3. *Suppose that $\{X_n\}$ is independent and that $A \in \mathcal{T} = \bigcap_{n=1}^{\infty} \sigma(X_n, X_{n+1}, \ldots)$. Then either $P(A) = 0$ or $P(A) = 1$.*

PROOF. Let $\mathcal{F}_0 = \bigcup_{k=1}^{\infty} \sigma(X_1, \ldots, X_k)$. The first thing to establish is that $\mathcal{F}_0$ is a field generating the σ-field $\sigma(X_1, X_2, \ldots)$. If B and C lie in $\mathcal{F}_0$, then $B \in \sigma(X_1, \ldots, X_j)$ and $C \in \sigma(X_1, \ldots, X_k)$ for some j and k; if $m = \max\{j, k\}$, then B and C both lie in $\sigma(X_1, \ldots, X_m)$, so that $B \cup C \in \sigma(X_1, \ldots, X_m) \subset \mathcal{F}_0$. Thus $\mathcal{F}_0$ is closed under the formation of finite unions; since it is similarly closed under complementation, $\mathcal{F}_0$ is a field. For $H \in \mathcal{R}^1$, $[X_n \in H] \in \mathcal{F}_0 \subset \sigma(\mathcal{F}_0)$, and hence X_n is measurable $\sigma(\mathcal{F}_0)$; thus $\mathcal{F}_0$ generates $\sigma(X_1, X_2, \ldots)$ (which in general is much larger than $\mathcal{F}_0$).

Suppose that A lies in $\mathcal{T}$. Then A lies in $\sigma(X_{k+1}, X_{k+2}, \ldots)$ for each k. Therefore, if $B \in \sigma(X_1, \ldots, X_k)$, then A and B are independent by Theorem 20.2. Therefore, A is independent of $\mathcal{F}_0$ and hence by Theorem 4.2 is also independent of $\sigma(X_1, X_2, \ldots)$. But then A is independent of itself: $P(A \cap A) = P(A)P(A)$. Therefore, $P(A) = P^2(A)$, which implies that $P(A)$ is either 0 or 1. ∎

As noted above, the set where $\sum_n X_n(\omega)$ converges satisfies the hypothesis of Theorem 22.3, and so does the set where $n^{-1}\sum_{k=1}^{n} X_k(\omega) \to 0$. In many similar cases it is very easy to prove by this theorem that a set at hand must have probability either 0 or 1. But to determine which of 0 and 1 is, in fact, the probability of the set may be extremely difficult.

Maximal Inequalities

Essential to the study of random series are maximal inequalities—inequalities concerning the maxima of partial sums. The best known is that of Kolmogorov.

Theorem 22.4. *Suppose that $X_1, \ldots, X_n$ are independent with mean 0 and finite variances. For $\alpha > 0$,*

$$(22.9) \qquad P\left[\max_{1 \le k \le n} |S_k| \ge \alpha\right] \le \frac{1}{\alpha^2} \mathrm{Var}[S_n].$$

PROOF. Let A_k be the set where $|S_k| \ge \alpha$ but $|S_j| < \alpha$ for $j < k$. Since the A_k are disjoint,

$$E[S_n^2] \ge \sum_{k=1}^{n} \int_{A_k} S_n^2 \, dP$$

$$= \sum_{k=1}^{n} \int_{A_k} \left[S_k^2 + 2S_k(S_n - S_k) + (S_n - S_k)^2 \right] dP$$

$$\ge \sum_{k=1}^{n} \int_{A_k} \left[S_k^2 + 2S_k(S_n - S_k) \right] dP.$$

Since A_k and S_k are measurable $\sigma(X_1, \ldots, X_k)$ and $S_n - S_k$ is measurable $\sigma(X_{k+1}, \ldots, X_n)$, and since the means are all 0, it follows by (21.19) and independence that $\int_{A_k} S_k(S_n - S_k) \, dP = 0$. Therefore,

$$E[S_n^2] \ge \sum_{k=1}^{n} \int_{A_k} S_k^2 \, dP \ge \sum_{k=1}^{n} \alpha^2 P(A_k)$$

$$= \alpha^2 P\left[\max_{1 \le k \le n} |S_k| \ge \alpha \right]. \qquad \blacksquare$$

By Chebyshev's inequality, $P[|S_n| \ge \alpha] \le \alpha^{-2} \operatorname{Var}[S_n]$. That this can be strengthened to (22.9) is an instance of a general phenomenon: For sums of independent variables, if $\max_{k \le n} |S_k|$ is large, then $|S_n|$ is probably large as well. Theorem 9.6 is an instance of this, and so is the following result, due to Etemadi.

Theorem 22.5. *Suppose that $X_1, \ldots, X_n$ are independent. For $\alpha \ge 0$,*

$$(22.10) \qquad P\left[\max_{1 \le k \le n} |S_k| \ge 3\alpha \right] \le 3 \max_{1 \le k \le n} P\{|S_k| \ge \alpha\}.$$

PROOF. Let B_k be the set where $|S_k| \ge 3\alpha$ but $|S_j| < 3\alpha$ for $j < k$. Since the B_k are disjoint,

$$P\left[\max_{1 \le k \le n} |S_k| \ge 3\alpha \right] \le P[|S_n| \ge \alpha] + \sum_{k=1}^{n-1} P(B_k \cap [|S_n| < \alpha])$$

$$\le P[|S_n| \ge \alpha] + \sum_{k=1}^{n-1} P(B_k \cap [|S_n - S_k| > 2\alpha])$$

$$= P\{|S_n| \ge \alpha\} + \sum_{k=1}^{n-1} P(B_k) P\{|S_n - S_k| > 2\alpha\}$$

$$\le P[|S_n| \ge \alpha] + \max_{1 \le k \le n} P[|S_n - S_k| \ge 2\alpha]$$

$$\le P[|S_n| \ge \alpha] + \max_{1 \le k \le n} (P[|S_n| \ge \alpha] + P\{|S_k| \ge \alpha\})$$

$$\le 3 \max_{1 \le k \le n} P\{|S_k| \ge \alpha\}. \qquad \blacksquare$$

If the X_k have mean 0 and Chebyshev's inequality is applied to the right side of (22.10), and if α is replaced by $\alpha/3$, the result is Kolmogorov's inequality (22.9) with an extra factor of 27 on the right side. For this reason, the two inequalities are equally useful for the applications in this section.

Convergence of Random Series

For independent X_n, the probability that ΣX_n converges is either 0 or 1. It is natural to try and characterize the two cases in terms of the distributions of the individual X_n.

Theorem 22.6.   *Suppose that $\{X_n\}$ is an independent sequence and $E[X_n] = 0$. If $\Sigma \operatorname{Var}[X_n] < \infty$, then ΣX_n converges with probability 1.*

PROOF.   By (22.9),

$$P\left[\max_{1 \le k \le r} |S_{n+k} - S_n| > \epsilon \right] \le \frac{1}{\epsilon^2} \sum_{k=1}^{r} \operatorname{Var}[X_{n+k}].$$

Since the sets on the left are nondecreasing in r, letting $r \to \infty$ gives

$$P\left[\sup_{k \ge 1} |S_{n+k} - S_n| > \epsilon \right] \le \frac{1}{\epsilon^2} \sum_{k=1}^{\infty} \operatorname{Var}[X_{n+k}].$$

Since $\Sigma \operatorname{Var}[X_n]$ converges,

(22.11) $$\lim_n P\left[\sup_{k \ge 1} |S_{n+k} - S_n| > \epsilon \right] = 0$$

for each ϵ.

Let $E(n, \epsilon)$ be the set where $\sup_{j,k \ge n} |S_j - S_k| > 2\epsilon$, and put $E(\epsilon) = \bigcap_n E(n, \epsilon)$. Then $E(n, \epsilon) \downarrow E(\epsilon)$, and (22.11) implies $P(E(\epsilon)) = 0$. Now $\bigcup_\epsilon E(\epsilon)$, where the union extends over positive rational ϵ, contains the set where the sequence $\{S_n\}$ is not fundamental (does not have the Cauchy property), and this set therefore has probability 0.   ∎

Example 22.2.   Let $X_n(\omega) = r_n(\omega) a_n$, where the r_n are the Rademacher functions on the unit interval—see (1.13). Then X_n has variance a_n^2, and so $\Sigma a_n^2 < \infty$ implies that $\Sigma r_n(\omega) a_n$ converges with probability 1. An interesting special case is $a_n = n^{-1}$. If the signs in $\Sigma \pm n^{-1}$ are chosen on the toss of a coin, then the series converges with probability 1. The alternating harmonic series $1 - 2^{-1} + 3^{-1} + \cdots$ is thus typical in this respect.   ∎

If ΣX_n converges with probability 1, then S_n converges with probability 1 to some finite random variable S. By Theorem 20.5, this implies that

$S_n \to_p S$. The reverse implication of course does not hold in general, but it does if the summands are independent.

Theorem 22.7. *For an independent sequence $\{X_n\}$, the S_n converge with probability 1 if and only if they converge in probability.*

PROOF.   It is enough to show that if $S_n \to_p S$, then $\{S_n\}$ is fundamental with probability 1. Since

$$P\Big[|S_{n+j} - S_n| \ge \epsilon\Big] \le P\Big[|S_{n+j} - S| \ge \frac{\epsilon}{2}\Big] + P\Big[|S_n - S| \ge \frac{\epsilon}{2}\Big],$$

$S_n \to_p S$ implies

(22.12) $$\lim_n \sup_{j \ge 1} P\Big[|S_{n+j} - S_n| \ge \epsilon\Big] = 0.$$

But by (22.10),

$$P\Big[\max_{1 \le j \le k} |S_{n+j} - S_n| \ge \epsilon\Big] \le 3 \max_{1 \le j \le k} P\Big[|S_{n+j} - S_n| \ge \frac{\epsilon}{3}\Big],$$

and therefore

$$P\Big[\sup_{k \ge 1} |S_{n+k} - S_n| > \epsilon\Big] \le 3 \sup_{k \ge 1} P\Big[|S_{n+k} - S_n| \ge \frac{\epsilon}{3}\Big].$$

It now follows by (22.12) that (22.11) holds, and the proof is completed as before.   ■

The final result in this direction, the *three-series theorem*, provides necessary and sufficient conditions for the convergence of $\sum X_n$ in terms of the individual distributions of the X_n. Let $X_n^{(c)}$ be X_n truncated at c: $X_n^{(c)} = X_n I_{[|X_n| \le c]}$.

Theorem 22.8. *Suppose that $\{X_n\}$ is independent, and consider the three series*

(22.13) $$\sum P[|X_n| > c], \qquad \sum E\big[X_n^{(c)}\big], \qquad \sum \mathrm{Var}\big[X_n^{(c)}\big].$$

In order that $\sum X_n$ converge with probability 1 it is necessary that the three series converge for all positive c and sufficient that they converge for some positive c.

PROOF OF SUFFICIENCY. Suppose that the series (22.13) converge, and put $m_n^{(c)} = E[X_n^{(c)}]$. By Theorem 22.6, $\Sigma(X_n^{(c)} - m_n^{(c)})$ converges with probability 1, and since $\Sigma m_n^{(c)}$ converges, so does $\Sigma X_n^{(c)}$. Since $P[X_n \neq X_n^{(c)} \text{ i.o.}] = 0$ by the first Borel–Cantelli lemma, it follows finally that ΣX_n converges with probability 1.   ■

Although it is possible to prove necessity in the three-series theorem by the methods of the present section, the simplest and clearest argument uses the central limit theorem as treated in Section 27. This involves no circularity of reasoning, since the three-series theorem is nowhere used in what follows.

PROOF OF NECESSITY. Suppose that ΣX_n converges with probability 1, and fix $c > 0$. Since $X_n \to 0$ with probability 1, it follows that $\Sigma X_n^{(c)}$ converges with probability 1 and, by the second Borel–Cantelli lemma, that $\Sigma P[|X_n| > c] < \infty$.

Let $M_n^{(c)}$ and $s_n^{(c)}$ be the mean and standard deviation of $S_n^{(c)} = \Sigma_{k=1}^{n} X_k^{(c)}$. If $s_n^{(c)} \to \infty$, then since the $X_n^{(c)} - m_n^{(c)}$ are uniformly bounded, it follows by the central limit theorem (see Example 27.4) that

$$(22.14) \qquad \lim_n P\left[x < \frac{S_n^{(c)} - M_n^{(c)}}{s_n^{(c)}} \leq y \right] = \frac{1}{\sqrt{2\pi}} \int_x^{y} e^{-t^2/2} \, dt.$$

And since $\Sigma X_n^{(c)}$ converges with probability 1, $s_n^{(c)} \to \infty$ also implies $S_n^{(c)}/s_n^{(c)} \to 0$ with probability 1, so that (Theorem 20.5)

$$(22.15) \qquad \lim_n P\left[|S_n^{(c)}/s_n^{(c)}| \geq \epsilon \right] = 0.$$

But (22.14) and (22.15) stand in contradiction: Since

$$P\left[x < \frac{S_n^{(c)} - M_n^{(c)}}{s_n^{(c)}} \leq y, \ \left| \frac{S_n^{(c)}}{s_n^{(c)}} \right| < \epsilon \right]$$

is greater than or equal to the probability in (22.14) minus that in (22.15), it is positive for all sufficiently large n (if $x < y$). But then

$$x - \epsilon < -M_n^{(c)}/s_n^{(c)} < y + \epsilon,$$

and this cannot hold simultaneously for, say, $(x - \epsilon, y + \epsilon) = (-1, 0)$ and $(x - \epsilon, y + \epsilon) = (0, 1)$. Thus $s_n^{(c)}$ cannot go to ∞, and the third series in (22.13) converges.

And now it follows by Theorem 22.6 that $\Sigma(X_n^{(c)} - m_n^{(c)})$ converges with probability 1, so that the middle series in (22.13) converges as well.   ■

Example 22.3. If $X_n = r_n a_n$, where r_n are the Rademacher functions, then $\Sigma a_n^2 < \infty$ implies that ΣX_n converges with probability 1. If ΣX_n converges, then a_n is bounded, and for large c the convergence of the third

series in (22.13) implies $\Sigma a_n^2 < \infty$: If the signs in $\Sigma \pm a_n$ are chosen on the toss of a coin, then the series converges with probability 1 or 0 according as Σa_n^2 converges or diverges. If Σa_n^2 converges but $\Sigma|a_n|$ diverges, then $\Sigma \pm a_n$ is with probability 1 conditionally but not absolutely convergent. ∎

Example 22.4. If $a_n \downarrow 0$ but $\Sigma a_n^2 = \infty$, then $\Sigma \pm a_n$ converges if the signs are strictly alternating, but diverges with probability 1 if they are chosen on the toss of a coin. ∎

Theorems 22.6, 22.7, and 22.8 concern conditional convergence, and in the most interesting cases, ΣX_n converges not because the X_n go to 0 at a high rate but because they tend to cancel each other out. In Example 22.4, the terms cancel well enough for convergence if the signs are strictly alternating, but not if they are chosen on the toss of a coin.

Random Taylor Series*

Consider a power series $\Sigma \pm z^n$, where the signs are chosen on the toss of a coin. The radius of convergence being 1, the series represents an analytic function in the open unit disk $D_0 = [z: |z| < 1]$ in the complex plane. The question arises whether this function can be extended analytically beyond D_0. The answer is no: With probability 1 the unit circle is the natural boundary.

Theorem 22.9. *Let* $\{X_n\}$ *be an independent sequence such that*

$$(22.16) \qquad P[X_n = 1] = P[X_n = -1] = \tfrac{1}{2}, \qquad n = 0, 1, \ldots .$$

There is probability 0 *that*

$$(22.17) \qquad F(\omega, z) = \sum_{n=0}^{\infty} X_n(\omega) z^n$$

coincides in D_0 *with a function analytic in an open set properly containing* D_0.

It will be seen in the course of the proof that the ω-set in question lies in $\sigma(X_0, X_1, \ldots)$ and hence has a probability. It is intuitively clear that if the set is measurable at all, it must depend only on the X_n for large n and hence must have probability either 0 or 1.

PROOF. Since

$$(22.18) \qquad |X_n(\omega)| = 1, \qquad n = 0, 1, \ldots .$$

---

*This topic, which requires complex variable theory, may be omitted.

with probability 1, the series in (22.17) has radius of convergence 1 outside a set of measure 0.

Consider an open disk $D = [z: |z - \zeta| < r]$, where $\zeta \in D_0$ and $r > 0$. Now (22.17) coincides in D_0 with a function analytic in $D_0 \cup D$ if and only if its expansion

$$F(\omega, z) = \sum_{m=0}^{\infty} \frac{1}{m!} F^{(m)}(\omega, \zeta)(z - \zeta)^m$$

about ζ converges at least for $|z - \zeta| < r$. Let A_D be the set of ω for which this holds. The coefficient

$$a_m(\omega) = \frac{1}{m!} F^{(m)}(\omega, \zeta) = \sum_{n=m}^{\infty} \binom{n}{m} X_n(\omega) \zeta^{n-m}$$

is a complex-valued random variable measurable $\sigma(X_m, X_{m+1}, \ldots)$. By the root test, $\omega \in A_D$ if and only if $\limsup_m |a_m(\omega)|^{1/m} \le r^{-1}$. For each m_0, the condition for $\omega \in A_D$ can thus be expressed in terms of $a_{m_0}(\omega), a_{m_0+1}(\omega), \ldots$ alone, and so $A_D \in \sigma(X_{m_0}, X_{m_0+1}, \ldots)$. Thus A_D has a probability, and in fact $P(A_D)$ is 0 or 1 by the zero-one law.

Of course, $P(A_D) = 1$ if $D \subset D_0$. The central step in the proof is to show that $P(A_D) = 0$ if D contains points not in D_0. Assume on the contrary that $P(A_D) = 1$ for such a D. Consider that part of the circumference of the unit circle that lies in D, and let k be an integer large enough that this arc has length exceeding $2\pi/k$. Define

$$Y_n(\omega) = \begin{cases} X_n(\omega) & \text{if } n \not\equiv 0 \ (\text{mod } k), \\ -X_n(\omega) & \text{if } n \equiv 0 \ (\text{mod } k). \end{cases}$$

Let B_D be the ω-set where the function

$$(22.19) \qquad G(\omega, z) = \sum_{n=0}^{\infty} Y_n(\omega) z^n$$

coincides in D_0 with a function analytic in $D_0 \cup D$.

The sequence $\{Y_0, Y_1, \ldots\}$ has the same structure as the original sequence: the Y_n are independent and assume the values ± 1 with probability $\frac{1}{2}$ each. Since B_D is defined in terms of the Y_n in the same way as A_D is defined in terms of the X_n, it is intuitively clear that $P(B_D)$ and $P(A_D)$ must be the same. Assume for the moment the truth of this statement, which is somewhat more obvious than its proof.

If for a particular ω each of (22.17) and (22.19) coincides in D_0 with a function analytic in $D_0 \cup D$, the same must be true of

$$(22.20) \qquad F(\omega, z) - G(\omega, z) = 2 \sum_{m=0}^{\infty} X_{mk}(\omega) z^{mk}.$$

Let $D_l = [ze^{2\pi i l/k}: z \in D]$. Since replacing z by $ze^{2\pi i/k}$ leaves the function (22.20) unchanged, it can be extended analytically to each $D_0 \cup D_l$, $l = 1, 2, \ldots$. Because of the choice of k, it can therefore be extended analytically to $[z: |z| < 1 + \epsilon]$ for some positive ϵ; but this is impossible if (22.18) holds, since the radius of convergence must then be 1.

Therefore, $A_D \cap B_D$ cannot contain a point ω satisfying (22.18). Since (22.18) holds with probability 1, this rules out the possibility $P(A_D) = P(B_D) = 1$ and by the zero–one law leaves only the possibility $P(A_D) = P(B_D) = 0$. Let A be the ω-set where (22.17) extends to a function analytic in some open set larger than D_0. Then $\omega \in A$ if and only if (22.17) extends to $D_0 \cup D$ for some $D = [z: |z - \zeta| < r]$ for which $D - D_0 \neq \varnothing$, r is rational, and ζ has rational real and imaginary parts; in other words, A is the countable union of A_D for such D. Therefore, A lies in $\sigma(X_0, X_1, \ldots)$ and has probability 0.

It remains only to show that $P(A_D) = P(B_D)$, and this is most easily done by comparing $\{X_n\}$ and $\{Y_n\}$ with a canonical sequence having the same structure. Put $Z_n(\omega) = (X_n(\omega) + 1)/2$, and let $T\omega$ be $\sum_{n=0}^{\infty} Z_n(\omega) 2^{-n-1}$ on the ω-set A^* where this sum lies in $(0, 1]$; on $\Omega - A^*$ let $T\omega$ be 1, say. Because of (22.16) $P(A^*) = 1$. Let $\mathscr{F} = \sigma(X_0, X_1, \ldots)$ and let $\mathscr{B}$ be the σ-field of Borel subsets of $(0, 1]$; then $T: \Omega \to (0, 1]$ is measurable $\mathscr{F}/\mathscr{B}$. Let $r_n(x)$ be the nth Rademacher function. If $M = [x: r_i(x) = u_i, i = 1, \ldots, n]$, where $u_i = \pm 1$ for each i, then $P(T^{-1}M) = P[\omega: X_i(\omega) = u_i, i = 0, 1, \ldots, n - 1] = 2^{-n}$, which is the Lebesgue measure $\lambda(M)$ of M. Since these sets form a π-system generating $\mathscr{B}$, $P(T^{-1}M) = \lambda(M)$ for all M in $\mathscr{B}$ (Theorem 3.3).

Let M_D be the set of x for which $\sum_{n=0}^{\infty} r_{n+1}(x) z^n$ extends analytically to $D_0 \cup D$. Then M_D lies in $\mathscr{B}$, this being a special case of the fact that A_D lies in $\mathscr{F}$. Moreover, if $\omega \in A^*$, then $\omega \in A_D$ if and only if $T\omega \in M_D$: $A^* \cap A_D = A^* \cap T^{-1}M_D$. Since $P(A^*) = 1$, it follows that $P(A_D) = \lambda(M_D)$.

This argument only uses (22.16), and therefore it applies to $\{Y_n\}$ and B_D as well. Therefore, $P(B_D) = \lambda(M_D) = P(A_D)$.   ∎

PROBLEMS

22.1. Suppose that $X_1, X_2, \ldots$ is an independent sequence and Y is measurable $\sigma(X_n, X_{n+1}, \ldots)$ for each n. Show that there exists a constant a such that $P[Y = a] = 1$.

22.2. Assume $\{X_n\}$ independent, and define $X_n^{(c)}$ as in Theorem 22.8. Prove that for $\Sigma |X_n|$ to converge with probability 1 it is necessary that $\Sigma P[|X_n| > c]$ and $\Sigma E[|X_n^{(c)}|]$ converge for all positive c and sufficient that they converge for some positive c. If the three series (22.13) converge but $\Sigma E[|X_n^{(c)}|] = \infty$, then there is probability 1 that ΣX_n converges conditionally but not absolutely.

22.3. ↑   (a) Generalize the Borel–Cantelli lemmas: Suppose X_n are nonnegative. If $\Sigma E[X_n] < \infty$, then ΣX_n converges with probability 1. If the X_n are independent and uniformly bounded, and if $\Sigma E[X_n] = \infty$, then ΣX_n diverges with probability 1.

(b) Construct independent, nonnegative X_n such that ΣX_n converges with probability 1 but $\Sigma E[X_n]$ diverges For an extreme example, arrange that $P[X_n > 0 \text{ i.o.}] = 0$ but $E[X_n] \equiv \infty$.

22.4. Show under the hypothesis of Theorem 22.6 that ΣX_n has finite variance and extend Theorem 22.4 to infinite sequences.

22.5. 20.14   22.1 ↑   Suppose that $X_1, X_2, \ldots$ are independent, each with the Cauchy distribution (20.45) for a common value of u.

(a) Show that $n^{-1}\Sigma_{k=1}^{n} X_k$ does not converge with probability 1. Contrast with Theorem 22.1.

(b) Show that $P[n^{-1} \max_{k \le n} X_k \le x] \to e^{-u/\pi x}$ for $x > 0$. Relate to Theorem 14.3.

22.6. If $X_1, X_2, \ldots$ are independent and identically distributed, and if $P[X_1 \ge 0] = 1$ and $P[X_1 > 0] > 0$, then $\Sigma_n X_n = \infty$ with probability 1. Deduce this from Theorem 22.1 and its corollary and also directly: find a positive ϵ such that $X_r > \epsilon$ infinitely often with probability 1.

22.7. Suppose that $X_1, X_2, \ldots$ are independent and identically distributed and $E[|X_1|] = \infty$. Use (21.9) to show that $\Sigma_n P[|X_n| \ge an] = \infty$ for each a, and conclude that $\sup_n n^{-1}|X_n| = \infty$ with probability 1. Now show that $\sup_n n^{-1}|S_n| = \infty$ with probability 1. Compare with the corollary to Theorem 22.1.

22.8. *Wald's equation.* Let $X_1, X_2, \ldots$ be independent and identically distributed with finite mean, and put $S_n = X_1 + \cdots + X_n$. Suppose that τ is a stopping time: τ has positive integers as values and $[\tau = n] \in \sigma(X_1, \ldots, X_n)$; see Section 7 for examples. Suppose also that $E[\tau] < \infty$.

(a) Prove that

(22.21) $E[S_\tau] = E[X_1]E[\tau].$

(b) Suppose that X_n is ± 1 with probabilities p and q, $p \ne q$, let τ be the first n for which S_n is $-a$ or b (a and b positive integers), and calculate $E[\tau]$. This gives the expected duration of the game in the gambler's ruin problem for unequal p and q.

22.9. 20.9 ↑   Let Z_n be 1 or 0 according as at time n there is or is not a record in the sense of Problem 20.9. Let $R_n = Z_1 + \cdots + Z_n$ be the number of records up to time n. Show that $R_n/\log n \to_P 1$.

22.10. 22.1 ↑   (a) Show that for an independent sequence $\{X_n\}$ the radius of convergence of the random Taylor series $\Sigma_n X_n z^n$ is r with probability 1 for some nonrandom r.

(b) Suppose that the X_n have the same distribution and $P[X_1 \ne 0] > 0$. Show that r is 1 or 0 according as $\log^+|X_1|$ has finite mean or not.

22.11. Suppose that $X_0, X_1, \ldots$ are independent and each is uniformly distributed over $[0, 2\pi]$. Show that with probability 1 the series $\Sigma_n e^{iX_n}z^n$ has the unit circle as its natural boundary.

22.12. Prove (what is essentially Kolmogorov's zero-one law) that if A is independent of a π-system $\mathscr{P}$ and $A \in \sigma(\mathscr{P})$, then $P(A)$ is either 0 or 1.

22.13. Suppose that $\mathscr{A}$ is a semiring containing Ω.

(a) Show that if $P(A \cap B) \le bP(B)$ for all $B \in \mathscr{A}$, and if $b < 1$ and $A \in \sigma(\mathscr{A})$, then $P(A) = 0$.

(b) Show that if $P(A \cap B) \le P(A)P(B)$ for all $B \in \mathscr{A}$, and if $A \in \sigma(\mathscr{A})$, then $P(A)$ is 0 or 1.

(c) Show that if $aP(B) \le P(A \cap B)$ for all $B \in \mathscr{A}$, and if $a > 0$ and $A \in \sigma(\mathscr{A})$, then $P(A) = 1$.

(d) Show that if $P(A)P(B) \le P(A \cap B)$ for all $B \in \mathscr{A}$, and if $A \in \sigma(\mathscr{A})$, then $P(A)$ is 0 or 1.

(e) Reconsider Problem 3.20

22.14. 22.12↑ *Burstin's theorem.* Let f be a Borel function on $[0, 1]$ with arbitrarily small periods: For each ϵ there is a p such that $0 < p < \epsilon$ and $f(x) = f(x + p)$ for $0 \le x \le 1 - p$. Show that such an f is constant almost everywhere:

(a) Show that it is enough to prove that $P(f^{-1}B)$ is 0 or 1 for every Borel set B, where P is Lebesgue measure on the unit interval.

(b) Show that $f^{-1}B$ is independent of each interval $[0, x]$, and conclude that $P(f^{-1}B)$ is 0 or 1.

(c) Show by example that f need not be constant.

22.15. Assume that $X_1, \ldots, X_n$ are independent and s, t, α are nonnegative Let

$$L(s) = \max_{k \le n} P[|S_k| \ge s], \quad R(s) = \max_{k \le n} P[|S_n - S_k| > s],$$

$$M(s) = P\left[\max_{k \le n} |S_k| \ge s\right], \quad T(s) = P[|S_n| \ge s].$$

(a) Following the first part of the proof of (22.10), show that

$$(22.22) \qquad M(s + t) \le T(t) + M(s + t)R(s).$$

(b) Take $s = 2\alpha$ and $t = \alpha$; use (22.22), together with the inequalities $T(s) \le L(s)$ and $R(2s) \le 2L(s)$, to prove Etemadi's inequality (22.10) in the form

$$(22.23) \qquad M(3\alpha) \le B_E(\alpha) = 1 \wedge 3L(\alpha).$$

(c) Carry the rightmost term in (22.22) to the left side, take $s = t = \alpha$, and prove *Ottaviani's* inequality:

$$(22.24) \qquad M(2\alpha) \le B_O(\alpha) = 1 \wedge \frac{T(\alpha)}{1 - R(\alpha)}.$$

(d) Prove

$$B_E(\alpha) \le 3B_O(\alpha/2), \quad B_O(\alpha) \le 3B_E(\alpha/6).$$

This shows that the Etemadi and Ottaviani inequalities are of the same power for most purposes (as, for example, for the proofs of Theorem 22.7 and (37.9)). Etemadi's inequality seems the more natural of the two. Neither inequality can replace (9.39) in the proof of the law of the iterated logarithm.

SECTION 23. THE POISSON PROCESS

Characterization of the Exponential Distribution

Suppose that X has the exponential distribution with parameter α:

$$(23.1) \qquad P[X > x] = e^{-\alpha x}, \qquad x \geq 0.$$

The definition (4.1) of conditional probability then gives

$$(23.2) \qquad P[X > x + y \mid X > x] = P[X > y], \qquad x, y \geq 0.$$

Image X as the waiting time for the occurrence of some event such as the arrival of the next customer at a queue or telephone call at an exchange. As observed in Section 14 (see (14.6)), (23.2) attributes to the waiting-time mechanism a lack of memory or aftereffect. And as shown in Section 14, the condition (23.2) implies that X has the distribution (23.1) for some positive α. Thus if in the sense of (23.2) there is no aftereffect in the waiting-time mechanism, then the waiting time itself necessarily follows the exponential law.

The Poisson Process

Consider next a stream or sequence of events, say arrivals of calls at an exchange. Let X_1 be the waiting time to the first event, let X_2 be the waiting time between the first and second events, and so on. The formal model consists of an infinite sequence $X_1, X_2, \ldots$ of random variables on some probability space, and $S_n = X_1 + \cdots + X_n$ represents the time of occurrence of the nth event; it is convenient to write $S_0 = 0$. The stream of events itself remains intuitive and unformalized, and the mathematical definitions and arguments are framed in terms of the X_n.

If no two of the events are to occur simultaneously, the S_n must be strictly increasing, and if only finitely many of the events are to occur in each finite interval of time, S_n must go to infinity:

$$(23.3) \qquad 0 = S_0(\omega) < S_1(\omega) < S_2(\omega) < \cdots, \qquad \sup_n S_n(\omega) = \infty.$$

This condition is the same thing as

$$(23.4) \qquad X_1(\omega) > 0, \quad X_2(\omega) > 0, \ldots, \qquad \sum_n X_n(\omega) = \infty.$$

Throughout the section it will be assumed that these conditions hold everywhere—for every ω. If they hold only on a set A of probability 1, and if $X_n(\omega)$ is redefined as $X_n(\omega) = 1$, say, for $\omega \notin A$, then the conditions hold everywhere and the joint distributions of the X_n and S_n are unaffected.

Condition $0°$. *For each* ω, (23.3) *and* (23.4) *hold*.

The arguments go through under the weaker condition that (23.3) and (23.4) hold with probability 1, but they then involve some fussy and uninteresting details. There are at the outset no further restrictions on the X_j; they are not assumed independent, for example, or identically distributed.

The number N_t of events that occur in the time interval $[0, t]$ is the largest integer n such that $S_n \le t$:

(23.5) $$N_t = \max[\, n : S_n \le t\,].$$

Note that $N_t = 0$ if $t < S_1 = X_1$; in particular, $N_0 = 0$. The number of events in $(s, t]$ is the increment $N_t - N_s$.
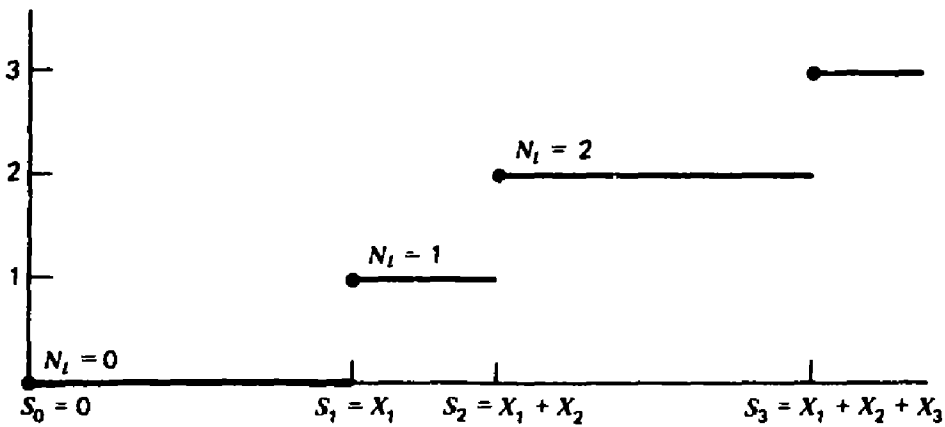
From (23.5) follows the basic relation connecting the N_t with the S_n:

(23.6) $$[\, N_t \ge n\,] = [\, S_n \le t\,].$$

From this follows

(23.7) $$[\, N_t = n\,] = [\, S_n \le t < S_{n+1}\,].$$

Each N_t is thus a random variable.

The collection $[N_t : t \ge 0]$ is a *stochastic process*, that is, a collection of random variables indexed by a parameter regarded as time. Condition $0°$ can be restated in terms of this process:

Condition $0°$. *For each* ω, $N_t(\omega)$ *is a nonnegative integer for* $t \ge 0$, $N_0(\omega) = 0$, *and* $\lim_{t \to \infty} N_t(\omega) = \infty$; *further, for each* ω, $N_t(\omega)$ *as a function of* t *is nondecreasing and right-continuous, and at the points of discontinuity the saltus* $N_t(\omega) - \sup_{s < t} N_s(\omega)$ *is exactly* 1.

It is easy to see that (23.3) and (23.4) and the definition (23.5) give random variables N_t having these properties. On the other hand, if the stochastic

process $[N_t\colon\ t\geq 0]$ is given and does have these properties, and if random variables are defined by $S_n(\omega)=\inf[t\colon\ N_t(\omega)\geq n]$ and $X_n(\omega)=S_n(\omega)-S_{n-1}(\omega)$, then (23.3) and (23.4) hold, and the definition (23.5) gives back the original N_t. Therefore, anything that can be said about the X_n can be stated in terms of the N_t, and conversely. The points $S_1(\omega),S_2(\omega),\ldots$ of $(0,\infty)$ are exactly the discontinuities of $N_t(\omega)$ as a function of t; because of the queueing example, it is natural to call them *arrival times*.

The program is to study the joint distributions of the N_t under conditions on the waiting times X_n and vice versa. The most common model specifies the independence of the waiting times and the absence of aftereffect:

Condition 1°. The X_n are independent, and each is exponentially distributed with parameter α.

In this case $P[X_n>0]=1$ for each n and $n^{-1}S_n\to\alpha^{-1}$ by the strong law of large numbers (Theorem 22.1), and so (23.3) and (23.4) hold with probability 1; to assume they hold everywhere (Condition 0°) is simply a convenient normalization.

Under Condition 1°, S_n has the distribution function specified by (20.40), so that $P[N_t\geq n]=\sum_{i=n}^{\infty}e^{-\alpha t}(\alpha t)^i/i!$ by (23.6), and

$$(23.8)\qquad P[N_t=n]=e^{-\alpha t}\frac{(\alpha t)^n}{n!}\,,\qquad n=0,1,\ldots.$$

Thus N_t has the Poisson distribution with mean αt. More will be proved in a moment.

Condition 2°. (i) *For $0<t_1<\cdots<t_k$ the increments $N_{t_1},N_{t_2}-N_{t_1},\ldots,N_{t_k}-N_{t_{k-1}}$ are independent.*

(ii) *The individual increments have the Poisson distribution:*

$$(23.9)\quad P[N_t-N_s=n]=e^{-\alpha(t-s)}\frac{(\alpha(t-s))^n}{n!}\,,\qquad n=0,1,\ldots,\quad 0\leq s<t.$$

Since $N_0=0$, (23.8) is a special case of (23.9). A collection $[N_t\colon\ t\geq 0]$ of random variables satisfying Condition 2° is called a *Poisson process*, and α is the *rate* of the process. As the increments are independent by (i), if $r<s<t$, then the distributions of N_s-N_r and N_t-N_s must convolve to that of N_t-N_r. But the requirement is consistent with (ii) because Poisson distributions with parameters u and v convolve to a Poisson distribution with parameter $u+v$.

Theorem 23.1. *Conditions 1° and 2° are equivalent in the presence of Condition 0°.*

PROOF OF $1° \rightarrow 2°$. Fix t, and consider the events that happen after time t. By (23.5), $S_{N_t} \leq t < S_{N_t + 1}$, and the waiting time from t to the first event following t is $S_{N_t + 1} - t$; the waiting time between the first and second events following t is $X_{N_t + 2}$; and so on. Thus

$$(23.10) \quad X_1^{(t)} = S_{N_t + 1} - t, \qquad X_2^{(t)} = X_{N_t + 2}, \qquad X_3^{(t)} = X_{N_t + 3}, \cdots$$

define the waiting times following t. By (23.6), $N_{t+s} - N_t \geq m$, or $N_{t+s} \geq N_t + m$, if and only if $S_{N_t + m} \leq t + s$, which is the same thing as $X_1^{(t)} + \cdots + X_m^{(t)} \leq s$. Thus

$$(23.11) \qquad N_{t+s} - N_t = \max\left[m : X_1^{(t)} + \cdots + X_m^{(t)} \leq s \right].$$

Hence $[N_{t+s} - N_t = m] = [X_1^{(t)} + \cdots + X_m^{(t)} \leq s < X_1^{(t)} + \cdots + X_{m+1}^{(t)}]$. A comparison of (23.11) and (23.5) shows that for fixed t the random variables $N_{t+s} - N_t$ for $s \geq 0$ are defined in terms of the sequence (23.10) in exactly the same way as the N_s are defined in terms of the original sequence of waiting times.

The idea now is to show that conditionally on the event $[N_t = n]$ the random variables (23.10) are independent and exponentially distributed. Because of the independence of the X_k and the basic property (23.2) of the exponential distribution, this seems intuitively clear. For a proof, apply (20.30). Suppose $y \geq 0$; if G_n is the distribution function of S_n, then since X_{n+1} has the exponential distribution,

$$P[S_n \leq t < S_{n+1}, S_{n+1} - t > y] = P[S_n \leq t, X_{n+1} > t + y - S_n]$$

$$= \int_{x \leq t} P[X_{n+1} > t + y - x]\, dG_n(x)$$

$$= e^{-\alpha y} \int_{x \leq t} P[X_{n+1} > t - x]\, dG_n(x)$$

$$= e^{-\alpha y} P[S_n \leq t, X_{n+1} > t - S_n].$$

By the assumed independence of the X_n,

$$P\left[S_{n+1} - t > y_1, X_{n+2} > y_2, \dots, X_{n+j} > y_j, S_n \leq t < S_{n+1}\right]$$

$$= P[S_{n+1} - t > y_1, S_n \leq t < S_{n+1}] e^{-\alpha y_2} \cdots e^{-\alpha y_j}$$

$$= P[S_n \leq t < S_{n+1}] e^{-\alpha y_1} \cdots e^{-\alpha y_j}.$$

If $H = (y_1, \infty) \times \cdots \times (y_j, \infty)$, this is

$$(23.12) \quad P\left[N_t = n, \left(X_1^{(t)}, \dots, X_j^{(t)}\right) \in H\right] = P[N_t = n] P\left[(X_1, \dots, X_j) \in H\right].$$

By Theorem 10.4, the equation extends from H of the special form above to all H in $\mathscr{R}^j$.

Now the event $[N_{s_i} = m_i, 1 \leq i \leq u]$ can be put in the form $[(X_1, \ldots, X_j) \in H]$, where $j = m_u + 1$ and H is the set of x in R^j for which $x_1 + \cdots + x_{m_i} \leq s_i < x_1 + \cdots + x_{m_i + 1}$, $1 \leq i \leq u$. But then $[(X_1^{(t)}, \ldots, X_j^{(t)}) \in H]$ is by (23.11) the same as the event $[N_{t+s_i} - N_t = m_i, 1 \leq i \leq u]$. Thus (23.12) gives

$$P\big[N_t = n, N_{t+s_i} - N_t = m_i, 1 \leq i \leq u\big] = P[N_t = n]P\big[N_{s_i} = m_i, 1 \leq i \leq u\big].$$

From this it follows by induction on k that if $0 = t_0 < t_1 < \cdots < t_k$, then

$$(23.13) \qquad P\big[N_{t_i} - N_{t_{i-1}} = n_i, 1 \leq i \leq k\big] = \prod_{i=1}^{k} P\big[N_{t_i - t_{i-1}} = n_i\big].$$

Thus Condition 1° implies (23.13) and, as already seen, (23.8). But from (23.13) and (23.8) follow the two parts of Condition 2°. ∎

PROOF OF 2° → 1°. If 2° holds, then by (23.6), $P[X_1 > t] = P[N_t = 0] = e^{-\alpha t}$, so that X_1 is exponentially distributed. To find the joint distribution of X_1 and X_2, suppose that $0 \leq s_1 < t_1 < s_2 < t_2$ and perform the calculation

$$P[s_1 < S_1 \leq t_1, s_2 < S_2 < t_2]$$

$$= P\big[N_{s_1} = 0, N_{t_1} - N_{s_1} = 1, N_{s_2} - N_{t_1} = 0, N_{t_2} - N_{s_2} \geq 1\big]$$

$$= e^{-\alpha s_1} \times \alpha(t_1 - s_1)e^{-\alpha(t_1 - s_1)} \times e^{-\alpha(s_2 - t_1)} \times (1 - e^{-\alpha(t_2 - s_2)})$$

$$= \alpha(t_1 - s_1)(e^{-\alpha s_2} - e^{-\alpha t_2}) = \iint_{\substack{s_1 < y_1 \leq t_1 \\ s_2 < y_2 \leq t_2}} \alpha^2 e^{-\alpha y_2} \, dy_1 \, dy_2.$$

Thus for a rectangle A contained in the open set $G = [(y_1, y_2): 0 < y_1 < y_2]$,

$$P[(S_1, S_2) \in A] = \int_A \alpha^2 e^{-\alpha y_2} \, dy_1 \, dy_2.$$

By inclusion–exclusion, this holds for finite unions of such rectangles and hence, by a passage to the limit, for countable ones. Therefore, it holds for $A = G \cap G'$ if G' is open. Since the open sets form a π-system generating the Borel sets, (S_1, S_2) has density $\alpha^2 e^{-\alpha y_2}$ on G (of course, the density is 0 outside G).

By a similar argument in R^k (the notation only is more complicated), $(S_1, \ldots, S_k)$ has density $\alpha^k e^{-\alpha y_k}$ on $[y: 0 < y_1 < \cdots < y_k]$. If a linear transformation $g(y) = x$ is defined by $x_i = y_i - y_{i-1}$, then $(X_1, \ldots; X_k) = g(S_1, \ldots, S_k)$ has by (20.20) the density $\prod_{i=1}^{k} \alpha e^{-\alpha x_i}$ (the Jacobian is identically 1). This proves Condition 1°. ∎

The Poisson Approximation

Other characterizations of the Poisson process depend on a generalization of the classical Poisson approximation to the binomial distribution.

Theorem 23.2. *Suppose that for each* n, $Z_{n1}, \ldots, Z_{nr_n}$ *are independent random variables and* Z_{nk} *assumes the values* 1 *and* 0 *with probabilities* p_{nk} *and* $1 - p_{nk}$. *If*

$$(23.14) \qquad \sum_{k=1}^{r_n} p_{nk} \to \lambda \geq 0, \qquad \max_{1 \leq k \leq r_n} p_{nk} \to 0,$$

then

$$(23.15) \qquad P\left[\sum_{k=1}^{r_n} Z_{nk} = i\right] \to e^{-\lambda} \frac{\lambda^i}{i!}, \qquad i = 0, 1, 2, \ldots .$$

If $\lambda = 0$, the limit in (23.15) is interpreted as 1 for $i = 0$ and 0 for $i \geq 1$. In the case where $r_n = n$ and $p_{nk} = \lambda/n$, (23.15) is the Poisson approximation to the binomial. Note that if $\lambda > 0$, then (23.14) implies $r_n \to \infty$.

PROOF. The argument depends on a construction like that in the proof of Theorem 20.4. Let $U_1, U_2, \ldots$ be independent random variables, each uniformly distributed over $[0, 1)$. For each p, $0 \leq p \leq 1$, split $[0, 1)$ into the two intervals $I_0(p) = [0, 1 - p)$ and $I_1(p) = [1 - p, 1)$, as well as into the sequence of intervals $J_i(p) = [\sum_{j=1}^{i-1} e^{-p} p^j / j!, \sum_{j=1}^{i} e^{-p} p^j / j!)$, $i = 0, 1, \ldots$. Define $V_{nk} = 1$ if $U_k \in I_1(p_{nk})$ and $V_{nk} = 0$ if $U_k \in I_0(p_{nk})$. Then $V_{n1}, \ldots, V_{nr_n}$ are independent, and V_{nk} assumes the values 1 and 0 with probabilities $P[U_k \in I_1(p_{nk})] = p_{nk}$ and $P[U_k \in I_0(p_{Nk})] = 1 - p_{nk}$. Since $V_{n1}, \ldots, V_{nr_n}$ have the same joint distribution as $Z_{n1}, \ldots, Z_{nr_n}$, (23.15) will follow if it is shown that $V_n = \sum_{k=1}^{r_n} V_{nk}$ satisfies

$$(23.16) \qquad P[V_n = i] \to e^{-\lambda} \frac{\lambda^i}{i!}.$$

Now define $W_{nk} = i$ if $U_k \in J_i(p_{nk})$, $i = 0, 1, \ldots$. Then $P[W_{nk} = i] = e^{-p_{nk}} p_{nk}^i / i!$—$W_{nk}$ has the Poisson distribution with mean p_{nk}. Since the W_{nk}

are independent, $W_n = \sum_{k=1}^{r_n} W_{nk}$ has the Poisson distribution with mean $\lambda_n = \sum_{k=1}^{r_n} p_{nk}$. Since $1 - p \le e^{-p}$, $J_1(p) \subset I_1(p)$ (see the diagram). Therefore,

$$P[V_{nk} \ne W_{nk}] = P[V_{nk} = 1 \ne W_{nk}] = P\left[U_k \in I_1(p_{nk}) - J_1(p_{nk})\right]$$

$$= p_{nk} - e^{-p_{nk}} p_{nk} \le p_{nk}^2,$$

and

$$P[V_n \ne W_n] \le \sum_{k=1}^{r_n} p_{nk}^2 \le \lambda_n \max_{1 \le k \le r_n} p_{nk} \to 0$$

by (23.14). And now (23.16) and (23.15) follow because

$$P[W_n = i] = e^{-\lambda_n} \lambda_n^i / i! \to e^{-\lambda} \lambda^i / i! \qquad \blacksquare$$

Other Characterizations of the Poisson Process

The condition (23.2) is an interesting characterization of the exponential distribution because it is essentially qualitative. There are qualitative characterizations of the Poisson process as well.

For each ω, the function $N_t(\omega)$ has a discontinuity at t if and only if $S_n(\omega) = t$ for some $n \ge 1$; t is a *fixed discontinuity* if the probability of this is positive. The condition that there be no fixed discontinuities is therefore

(23.17) $P[S_n = t] = 0, \qquad t \ge 0, \quad n \ge 1;$

that is, each of $S_1, S_2, \ldots$ has a continuous distribution function. Of course there is probability 1 (under Condition $0°$) that $N_t(\omega)$ has a discontinuity *somewhere* (and indeed has infinitely many of them). But (23.17) ensures that a t *specified in advance* has probability 0 of being a discontinuity, or time of an arrival. The Poisson process satisfies this natural condition.

Theorem 23.3. *If Condition $0°$ holds and $[N_t: t \ge 0]$ has independent increments and no fixed discontinuities, then each increment has a Poisson distribution.*

This is Prékopà's theorem. The conclusion is not that $[N_t: t \ge 0]$ is a Poisson process, because the mean of $N_t - N_s$ need not be proportional to $t - s$. If φ is an arbitrary nondecreasing, continuous function on $[0, \infty)$ and $\varphi(0) = 0$, and if $[N_t: t \ge 0]$ is a Poisson process, then $N_{\varphi(t)}$ satisfies the conditions of the theorem.[†]

PROOF. The problem is to show for $t' < t''$ that $N_{t''} - N_{t'}$ has for some $\lambda \ge 0$ a Poisson distribution with mean λ, a unit mass at 0 being regarded as a Poisson distribution with mean 0.

[†]This is in fact the general process satisfying them; see Problem 23.8

The procedure is to construct a sequence of partitions

$$(23.18) \qquad t' = t_{n0} < t_{n1} < \cdots < t_{nr_n} = t''$$

of $[t', t'']$ with three properties. First, each decomposition refines the preceding one: each t_{nk} is a $t_{n+1, j}$. Second,

$$(23.19) \qquad \sum_{k=1}^{r_n} P\big[N_{t_{nk}} - N_{t_{n,k-1}} \geq 1 \big] \uparrow \lambda$$

for some finite λ and

$$(23.20) \qquad \max_{1 \leq k \leq r_n} P\big[N_{t_{nk}} - N_{t_{n,k-1}} \geq 1 \big] \to 0.$$

Third,

$$(23.21) \qquad P\Big[\max_{1 \leq k \leq r_n} \big(N_{t_{nk}} - N_{t_{n,k-1}} \big) \geq 2 \Big] \to 0.$$

Once the partitions have been constructed, the rest of the proof is easy: Let Z_{nk} be 1 or 0 according as $N_{t_{nk}} - N_{t_{n,k-1}}$ is positive or not. Since $[N_t : t \geq 0]$ has independent increments, the Z_{nk} are independent for each n. By Theorem 23.2, therefore, (23.19) and (23.20) imply that $Z_n = \sum_{k=1}^{r_n} Z_{nk}$ satisfies $P[Z_n = i] \to e^{-\lambda} \lambda^i / i!$ Now $N_{t''} - N_{t'} \geq Z_n$, and there is strict inequality if and only if $N_{t_{nk}} - N_{t_{n,k-1}} \geq 2$ for some k. Thus (23.21) implies $P[N_{t''} - N_{t'} \neq Z_n] \to 0$, and therefore $P[N_{t''} - N_{t'} = i] = e^{-\lambda} \lambda^i / i!$

To construct the partitions, consider for each t the distance $D_t = \inf_{m \geq 1} |t - S_m|$ from t to the nearest arrival time. Since $S_m \to \infty$, the infimum is achieved. Further, $D_t = 0$ if and only if $S_m = t$ for some m, and since by hypothesis there are no fixed discontinuities, the probability of this is 0: $P[D_t = 0] = 0$. Choose δ_t so that $0 < \delta_t < n^{-1}$ and $P[D_t \leq \delta_t] < n^{-1}$. The intervals $(t - \delta_t, t + \delta_t)$ for $t' \leq t \leq t''$ cover $[t', t'']$. Choose a finite subcover, and in (23.18) take the t_{nk} for $0 < k < r_n$ to be the endpoints (of intervals in the subcover) that are contained in (t', t''). By the construction,

$$(23.22) \qquad \max_{1 \leq k \leq r_n} \big(t_{nk} - t_{n,k-1} \big) \to 0,$$

and the probability that $(t_{n,k-1}, t_{nk}]$ contains some S_m is less than n^{-1}. This gives a sequence of partitions satisfying (23.20). Inserting more points in a partition cannot increase the maxima in (23.20) and (23.22), and so it can be arranged that each partition refines the preceding one.

To prove (23.21) it is enough (Theorem 4.1) to show that the limit superior of the sets involved has probability 0. It is in fact empty: If for infinitely many n, $N_{t_{nk}}(\omega) - N_{t_{n,k-1}}(\omega) \geq 2$ holds for some $k \leq r_n$, then by (23.22), $N_t(\omega)$ as a

function of t has in $[t', t'']$ discontinuity points (arrival times) arbitrarily close together, which requires $S_m(\omega) \in [t', t'']$ for infinitely many m, in violation of Condition $0°$.

It remains to prove (23.19). If Z_{nk} and Z_n are defined as above and $p_{nk} = P[Z_{nk} = 1]$, then the sum in (23.19) is $\sum_k p_{nk} = E[Z_n]$. Since $Z_{n+1} \geq Z_n$, $\sum_k p_{nk}$ is nondecreasing in n. Now

$$P[N_{t''} - N_{t'} = 0] = P[Z_{nk} = 0, \ k \leq r_n]$$

$$= \prod_{k=1}^{r_n} (1 - p_{nk}) \leq \exp\left[-\sum_{k=1}^{r_n} p_{nk}\right].$$

If the left-hand side here is positive, this puts an upper bound on $\sum_k p_{nk}$, and (23.19) follows. But suppose $P[N_{t''} - N_{t'} = 0] = 0$. If s is the midpoint of t' and t'', then since the increments are independent, one of $P[N_s - N_{t'} = 0]$ and $P[N_{t''} - N_s = 0]$ must vanish. It is therefore possible to find a nested sequence of intervals $[u_m, v_m]$ such that $v_m - u_m \to 0$ and the event $A_m = [N_{v_m} - N_{u_m} \geq 1]$ has probability 1. But then $P(\bigcap_m A_m) = 1$, and if t is the point common to the $[u_m, v_m]$, there is an arrival at t with probability 1, contrary to the assumption that there are no fixed discontinuities. ∎

Theorem 23.3 in some cases makes the Poisson model quite plausible. The increments will be essentially independent if the arrivals to time s cannot seriously deplete the population of potential arrivals, so that N_s has for $t > s$ negligible effect on $N_t - N_s$. And the condition that there are no fixed discontinuities is entirely natural. These conditions hold for arrivals of calls at a telephone exchange if the rate of calls is small in comparison with the population of subscribers and calls are not placed at fixed, predetermined times. If the arrival rate is essentially constant, this leads to the following condition.

Condition $3°$. (i) *For* $0 < t_1 < \cdots < t_k$ *the increments* $N_{t_1}, N_{t_2} - N_{t_1}, \ldots,$ $N_{t_k} - N_{t_{k-1}}$ *are independent.*

(ii) *The distribution of* $N_t - N_s$ *depends only on the difference* $t - s$.

Theorem 23.4. *Conditions* $1°$, $2°$, *and* $3°$ *are equivalent in the presence of Condition* $0°$.

PROOF. Obviously Condition $2°$ implies $3°$. Suppose that Condition $3°$ holds. If J_t is the saltus at t $(J_t = N_t - \sup_{s < t} N_s)$, then $[N_t - N_{t-n^{-1}} \geq 1] \downarrow$ $[J_t \geq 1]$, and it follows by (ii) of Condition $3°$ that $P[J_t \geq 1]$ is the same for all t. But if the value common to $P[J_t \geq 1]$ is positive, then by the independence of the increments and the second Borel–Cantelli lemma there is probability 1 that $J_t \geq 1$ for infinitely many rational t in $(0, 1)$, for example, which contradicts Condition $0°$.

By Theorem 23.3, then, the increments have Poisson distributions. If $f(t)$ is the mean of N_t, then $N_t - N_s$ for $s < t$ must have mean $f(t) - f(s)$ and must by (ii) have mean $f(t - s)$; thus $f(t) = f(s) + f(t - s)$. Therefore, f satisfies Cauchy's functional equation [A20] and, being nondecreasing, must have the form $f(t) = \alpha t$ for $\alpha \geq 0$. Condition $0°$ makes $\alpha = 0$ impossible. ∎

One standard way of deriving the Poisson process is by differential equations.

Condition 4°.   If $0 < t_1 < \cdots < t_k$ and if $n_1, \ldots, n_k$ are nonnegative integers, then

$$(23.23) \qquad P\left[N_{t_k + h} - N_{t_k} = 1 | N_{t_j} = n_j, j \leq k \right] = \alpha h + o(h)$$

and

$$(23.24) \qquad P\left[N_{t_k + h} - N_{t_k} \geq 2 | N_{t_j} = n_j, j \leq k \right] = o(h)$$

as $h \downarrow 0$. Moreover, $[N_t : t \geq 0]$ has no fixed discontinuities.

The occurrences of $o(h)$ in (23.23) and (23.24) denote functions, say $\phi_1(h)$, and $\phi_2(h)$, such that $h^{-1}\phi_i(h) \to 0$ as $h \downarrow 0$; the ϕ_i may depend a priori on $k, t_1, \ldots, t_k$, and $n_1, \ldots, n_k$ as well as on h. It is assumed in (23.23) and (23.24) that the conditioning events have positive probability, so that the conditional probabilities are well defined.

Theorem 23.5.   *Conditions $1°$ through $4°$ are all equivalent in the presence of Condition $0°$.*

PROOF OF $2° \to 4°$.   For a Poisson process with rate α, the left-hand sides of (23.23) and (23.24) are $e^{-\alpha h}\alpha h$ and $1 - e^{-\alpha h} - e^{-\alpha h}\alpha h$, and these are $\alpha h + o(h)$ and $o(h)$, respectively, because $e^{-\alpha h} = 1 - \alpha h + o(h)$. And by the argument in the preceding proof, the process has no fixed discontinuities. ∎

PROOF OF $4° \to 2°$.   Fix k, the t_j, and the n_j; denote by A the event $[N_{t_j} = n_j, j \leq k]$; and for $t \geq 0$ put $p_n(t) = P[N_{t_k + t} - N_{t_k} = n | A]$. It will be shown that

$$(23.25) \qquad p_n(t) = e^{-\alpha t}\frac{(\alpha t)^n}{n!}, \qquad n = 0, 1, \ldots .$$

This will also be proved for the case in which $p_n(t) = P[N_t = n]$. Condition 2° will then follow by induction.

If $t > 0$ and $|t - s| < n^{-1}$, then

$$\left| P[N_t = n] - P[N_s = n] \right| \le P[N_t \ne N_s] \le P[N_{t+n^{-1}} - N_{t-n^{-1}} \ge 1].$$

As $n \to \infty$, the right side here decreases to the probability of a discontinuity at t, which is 0 by hypothesis. Thus $P[N_t = n]$ is continuous at t. The same kind of argument works for conditional probabilities and for $t = 0$, and so $p_n(t)$ is continuous for $t \ge 0$.

To simplify the notation, put $D_t = N_{t_k+t} - N_{t_k}$. If $D_{t+h} = n$, then $D_t = m$ for some $m \le n$. If $t > 0$, then by the rules for conditional probabilities,

$$p_n(t + h) = p_n(t) P\left[D_{t+h} - D_t = 0 | A \cap [D_t = n] \right]$$

$$+ p_{n-1}(t) P\left[D_{t+h} - D_t = 1 | A \cap [D_t = n - 1] \right]$$

$$+ \sum_{m=0}^{n-2} p_m(t) P\left[D_{t+h} - D_t = n - m | A \cap [D_t = m] \right].$$

For $n \le 1$, the final sum is absent, and for $n = 0$, the middle term is absent as well. This holds in the case $p_n(t) = P[N_t = n]$ if $D_t = N_t$ and $A = \Omega$. (If $t = 0$, some of the conditioning events here are empty; hence the assumption $t > 0$.) By (23.24), the final sum is $o(h)$ for each fixed n. Applying (23.23) and (23.24) now leads to

$$p_n(t + h) = p_n(t)(1 - \alpha h) + p_{n-1}(t)\alpha h + o(h),$$

and letting $h \downarrow 0$ gives

(23.26) $$p_n'(t) = -\alpha p_n(t) + \alpha p_{n-1}(t).$$

In the case $n = 0$, take $p_{-1}(t)$ to be identically 0. In (23.26), $t > 0$ and $p_n'(t)$ is a right-hand derivative. But since $p_n(t)$ and the right side of the equation are continuous on $[0, \infty)$, (23.26) holds also for $t = 0$ and $p_n'(t)$ can be taken as a two-sided derivative for $t > 0$ [A22].

Now (23.26) gives [A23]

$$p_n(t) = e^{-\alpha t}\left[p_n(0) + \alpha \int_0^t p_{n-1}(s) e^{\alpha s}\, ds \right].$$

Since $p_n(0)$ is 1 or 0 as $n = 0$ or $n > 0$, (23.25) follows by induction on n. ■

Stochastic Processes

The Poisson process $[N_t: t \geq 0]$ is one example of a *stochastic process*—that is, a collection of random variables (on some probability space $(\Omega, \mathscr{F}, P)$) indexed by a parameter regarded as representing time. In the Poisson case, time is *continuous*. In some cases the time is *discrete*: Section 7 concerns the sequence $\{F_n\}$ of a gambler's fortunes; there n represents time, but time that increases in jumps.

Part of the structure of a stochastic process is specified by its *finite-dimensional distributions*. For any finite sequence $t_1, \ldots, t_k$ of time points, the k-dimensional random vector $(N_{t_1}, \ldots, N_{t_k})$ has a distribution $\mu_{t_1 \cdots t_k}$ over R^k. These measures $\mu_{t_1 \cdots t_k}$ are the finite-dimensional distributions of the process. Condition 2° of this section in effect specifies them for the Poisson case:

$$(23.27) \quad P\left[N_{t_j} = n_j, j \leq k\right] = \prod_{j=1}^{k} e^{-\alpha(t_j - t_{j-1})} \frac{\left(\alpha(t_j - t_{j-1})\right)^{n_j - n_{j-1}}}{(n_j - n_{j-1})!}$$

if $0 \leq n_1 \leq \cdots \leq n_k$ and $0 \leq t_1 < \cdots < t_k$ (take $n_0 = t_0 = 0$).

The finite-dimensional distributions do not, however, contain all the mathematically interesting information about the process in the case of continuous time. Because of (23.3), (23.4), and the definition (23.5), for each fixed ω, $N_t(\omega)$ as a function of t has the regularity properties given in the second version of Condition 0°. These properties are used in an essential way in the proofs.

Suppose that $f(t)$ is t or 0 according as t is rational or irrational. Let N_t be defined as before, and let

$$(23.28) \qquad\qquad M_t(\omega) = N_t(\omega) + f(t + X_1(\omega)).$$

If R is the set of rationals, then $P[\omega: f(t + X_1(\omega)) \neq 0] = P[\omega: X_1(\omega) \in R - t] = 0$ for each t because $R - t$ is countable and X_1 has a density. Thus $P[M_t = N_t] = 1$ for each t, and so the stochastic process $[M_t: t \geq 0]$ has the same finite-dimensional distributions as $[N_t: t \geq 0]$. For ω fixed, however, $M_t(\omega)$ as a function of t is everywhere discontinuous and is neither monotone nor exclusively integer-valued.

The functions obtained by fixing ω and letting t vary are called the *path functions* or *sample paths* of the process. The example above shows that the finite-dimensional distributions do not suffice to determine the character of the path functions. In specifying a stochastic process as a model for some phenomenon, it is natural to place conditions on the character of the sample paths as well as on the finite-dimensional distributions. Condition 0° was imposed throughout this section to ensure that the sample paths are nondecreasing, right-continuous, integer-valued step functions, a natural condition if N_t is to represent the number of events in $[0, t]$. Stochastic processes in continuous time are studied further in Chapter 7.

PROBLEMS

Assume the Poisson processes here satisfy Condition $0°$ as well as Condition $1°$.

23.1. Show that the minimum of independent exponential waiting times is again exponential and that the parameters add.   .

23.2. 20.17↑ Show that the time S_n of the nth event in a Poisson stream has the gamma density $f(x; \alpha, n)$ as defined by (20.47). This is sometimes called the *Erlang* density.

23.3. Let $A_t = t - S_{N_t}$ be the time back to the most recent event in the Poisson stream (or to 0), and let $B_t = S_{N_t+1} - t$ be the time forward to the next event. Show that A_t and B_t are independent, that B_t is distributed as X_1 (exponentially with parameter α), and that A_t is distributed as $\min\{X_1, t\}$: $P[A_t \le t]$ is 0, $1 - e^{-\alpha x}$, or 1 as $x < 0$, $0 \le x < t$, or $x \ge t$.

23.4. ↑ Let $L_t = A_t + B_t = S_{N_t+1} - S_{N_t}$ be the length of the interarrival interval covering t.
(a) Show that L_t has density

$$d_t(x) = \begin{cases} \alpha^2 x e^{-\alpha x} & \text{if } 0 < x < t, \\ \alpha(1 + \alpha t)e^{-\alpha x} & \text{if } x \ge t. \end{cases}$$

(b) Show that $E[L_t]$ converges to $2E[X_1]$ as $t \to \infty$. This seems paradoxical because L_t is one of the X_n. Give an intuitive resolution of the apparent paradox.

23.5. *Merging Poisson streams.* Define a process $\{N_t\}$ by (23.5) for a sequence $\{X_n\}$ of random variables satisfying (23.4). Let $\{X_n'\}$ be a second sequence of random variables, on the same probability space, satisfying (23.4), and define $\{N_t'\}$ by $N_t' = \max[n: X_1' + \cdots + X_n' \le t]$. Define $\{N_t''\}$ by $N_t'' = N_t + N_t'$. Show that, if $\sigma(X_1, X_2, \dots)$ and $\sigma(X_1', X_2', \dots)$ are independent and $\{N_t\}$ and $\{N_t'\}$ are Poisson processes with respective rates α and β, then $\{N_t''\}$ is a Poisson process with rate $\alpha + \beta$.

23.6. ↑ The nth and $(n+1)$st events in the process $\{N_t\}$ occur at times S_n and S_{n+1}.
(a) Find the distribution of the number $N_{S_{n+1}}' - N_{S_n}'$ of events in the other process during this time interval.
(b) Generalize to $N_{S_m}' - N_{S_n}'$.

23.7. Suppose that $X_1, X_2, \dots$ are independent and exponentially distributed with parameter α, so that (23.5) defines a Poisson process $\{N_t\}$. Suppose that $Y_1, Y_2, \dots$ are independent and identically distributed and that $\sigma(X_1, X_2, \dots)$ and $\sigma(Y_1, Y_2, \dots)$ are independent. Put $Z_t = \sum_{k \le N_t} Y_k$. This is the *compound Poisson process*. If, for example, the event at time S_n in the original process

represents an insurance claim, and if Y_n represents the amount of the claim, then Z_t represents the total claims to time t.

(a) If $Y_k = 1$ with probability 1, then $\{Z_t\}$ is an ordinary Poisson process.

(b) Show that $\{Z_t\}$ has independent increments and that $Z_{s+t} - Z_s$ has the same distribution as Z_t.

(c) Show that, if Y_k assumes the values 1 and 0 with probabilities p and $1 - p$ ($0 < p < 1$), then $\{Z_t\}$ is a Poisson process with rate $p\alpha$.

23.8. Suppose a process satisfies Condition $0°$ and has independent, Poisson-distributed increments and no fixed discontinuities. Show that it has the form $\{N_{\varphi(t)}\}$, where $\{N_t\}$ is a standard Poisson process and φ is a nondecreasing, continuous function on $[0, \infty)$ with $\varphi(0) = 0$.

23.9. If the waiting times X_n are independent and exponentially distributed with parameter α, then $S_n/n \to \alpha^{-1}$ with probability 1, by the strong law of large numbers. From $\lim_{t \to \infty} N_t = \infty$ and $S_{N_t} \leq t < S_{N_t + 1}$ deduce that $\lim_{t \to \infty} N_t/t = \alpha$ with probability 1.

23.10. ↑ (a) Suppose that $X_1, X_2, \ldots$ are positive, and assume directly that $S_n/n \to m$ with probability 1, as happens if the X_n are independent and identically distributed with mean m. Show that $\lim_t N_t/t = 1/m$ with probability 1.

(b) Suppose now that $S_n/n \to \infty$ with probability 1, as happens if the X_n are independent and identically distributed and have infinite mean. Show that $\lim_t N_t/t = 0$ with probability 1.

The results in Problem 23.10 are theorems in *renewal theory*: A component of some mechanism is replaced each time it fails or wears out. The X_n are the lifetimes of the successive components, and N_t is the number of replacements, or renewals, to time t.

23.11. 20.7 23.10↑ Consider a persistent, irreducible Markov chain, and for a fixed state j let N_n be the number of passages through j up to time n. Show that $N_n/n \to 1/m$ with probability 1, where $m = \sum_{k=1}^{\infty} k f_{jj}^{(k)}$ is the mean return time (replace $1/m$ by 0 if this mean is infinite). See Lemma 3 in Section 8.

23.12. Suppose that X and Y have Poisson distributions with parameters α and β. Show that $|P[X = i] - P[Y = i]| \leq |\alpha - \beta|$. *Hint:* Suppose that $\alpha < \beta$, and represent Y as $X + D$, where X and D are independent and have Poisson distributions with parameters α and $\beta - \alpha$.

23.13. ↑ Use the methods in the proof of Theorem 23.2 to show that the error in (23.15) is bounded uniformly in i by $|\lambda - \lambda_n| + \lambda_n \max_k p_{nk}$.

SECTION 24. THE ERGODIC THEOREM*

Even though chance necessarily involves the notion of change, the laws governing the change may themselves remain constant as time passes: If time

---

*This section may be omitted. There is more on ergodic theory in Section 36.

does not alter the roulette wheel, the gambler's fortunes fluctuate according to constant probability laws. The ergodic theorem is a version of the strong law of large numbers general enough to apply to any system governed by probability laws that are invariant in time.

Measure-Preserving Transformations

Let $(\Omega, \mathscr{F}, P)$ be a probability space. A mapping $T: \Omega \to \Omega$ is a *measure-preserving transformation* if it is measurable $\mathscr{F}/\mathscr{F}$ and $P(T^{-1}A) = P(A)$ for all A in $\mathscr{F}$, from which it follows that $P(T^{-n}A) = P(A)$ for $n \geq 0$. If, further, T is a one-to-one mapping onto Ω and the point mapping T^{-1} is measurable $\mathscr{F}/\mathscr{F}$ ($TA \in \mathscr{F}$ for $A \in \mathscr{F}$), then T is *invertible*; in this case T^{-1} automatically preserves measure: $P(A) = P(T^{-1}TA) = P(TA)$.

The first result is a simple consequence of the π–λ theorem (or Theorems 13.1(i) and 3.3).

Lemma 1. *If $\mathscr{P}$ is a π-system generating $\mathscr{F}$, and if $T^{-1}A \in \mathscr{F}$ and $P(T^{-1}A) = P(A)$ for $A \in \mathscr{P}$, then T is a measure-preserving transformation.*

Example 24.1. *The Bernoulli shift.* Let S be a finite set, and consider the space S^{∞} of sequences (2.15) of elements of S. Define the *shift* T by

$$(24.1) \qquad T\omega = (z_2(\omega), z_3(\omega), \ldots);$$

the first element of $\omega = (z_1(\omega), z_2(\omega), \ldots)$ is lost, and T shifts the remaining elements one place to the left: $z_k(T\omega) = z_{k+1}(\omega)$ for $k \geq 1$. If A is a cylinder (2.17), then

$$(24.2) \qquad T^{-1}A = \left[\omega: (z_2(\omega), \ldots, z_{n+1}(\omega)) \in H \right]$$

$$= \left[\omega: (z_1(\omega), \ldots, z_{n+1}(\omega)) \in S \times H \right]$$

is another cylinder, and since the cylinders generate the basic σ-field $\mathscr{C}$, T is measurable $\mathscr{C}/\mathscr{C}$.

For probabilities p_u on S (nonnegative and summing to 1), define product measure P on the field $\mathscr{C}_0$ of cylinders by (2.21). Then P is consistently defined and countably additive (Theorem 2.3) and hence extends to a probability measure on $\mathscr{C} = \sigma(\mathscr{C}_0)$. Since the thin cylinders (2.16) form a π-system that generates $\mathscr{C}$, P is completely determined by the probabilities it assigns to them:

$$(24.3) \qquad P\left[\omega: (z_1(\omega), \ldots, z_n(\omega)) = (u_1, \ldots, u_n) \right] = p_{u_1} \cdots p_{u_n}.$$

If A is the thin cylinder on the left here, then by (24.2),

$$(24.4) \qquad P(T^{-1}A) = \sum_{u \in S} p_u p_{u_1} \cdots p_{u_n} = p_{u_1} \cdots p_{u_n} = P(A),$$

and it follows by the lemma that T preserves P. This T is the *Bernoulli shift*.

If $A = [\omega: z_1(\omega) = u]$ $(u \in S)$, then $I_A(T^{k-1}\omega)$ is 1 or 0 according as $z_k(\omega)$ is u or not. Since by (24.3) the events $[\omega: z_k(\omega) = u]$ are independent, each with probability p_u, the random variables $I_A(T^{k-1}\omega)$ are independent and take the value 1 with probability $p_u = P(A)$. By the strong law of large numbers, therefore,

$$(24.5) \qquad \lim_n \frac{1}{n} \sum_{k=1}^{n} I_A(T^{k-1}\omega) = P(A)$$

with probability 1.   ∎

Example 24.1 gives a model for independent trials, and that T preserves P means the probability laws governing the trials are invariant in time. In the present section, it is this invariance of the probability laws that plays the fundamental role; independence is a side issue.

The *orbit* under T of the point ω is the sequence $(\omega, T\omega, T^2\omega, \ldots)$, and (24.5) can be expressed by saying that the orbit enters the set A with asymptotic relative frequency $P(A)$. For $A = [\omega: (z_1(\omega), z_2(\omega)) = (u_1, u_2)]$, the $I_A(T^{k-1}\omega)$ are not independent, but (24.5) holds anyway. In fact, for the Bernoulli shift, (24.5) holds with probability 1 *whatever* A may be $(A \in \mathcal{F})$. This is one of the consequences of the ergodic theorem (Theorem 24.1). What is more, according to this theorem the limit in (24.5) exists with probability 1 (although it may not be constant in ω) if T is an arbitrary measure-preserving transformation on an arbitrary probability space, of which there are many examples.

Example 24.2. *The Markov shift.* Let $P = [p_{ij}]$ be a stochastic matrix with rows and columns indexed by the finite set S, and let π_i be probabilities on S. Replace (2.21) by $P(A) = \sum_H \pi_{u_1} p_{u_1 u_2} \cdots p_{u_{n-1} u_n}$. The argument in Section 2 showing that product measure is consistently defined and finitely additive carries over to this new measure: since the rows of the transition matrix add to 1,

$$\sum_{u_{n+1}} \sum_{u_m} p_{u_n u_{n+1}} \cdots p_{u_{m-1} u_m} = 1,$$

and so the argument involving (2.23) goes through. The new measure is again

countably additive on $\mathscr{C}_0$ (Theorem 2.3) and so extends to $\mathscr{C}$. This probability measure P on $\mathscr{C}$ is uniquely determined by the condition

$$(24.6) \quad P\big[\omega: (z_1(\omega), \ldots, z_n(\omega)) = (u_1, \ldots, u_n)\big] = \pi_{u_1} p_{u_1 u_2} \cdots p_{u_{n-1} u_n}.$$

Thus the coordinate functions $z_n(\cdot)$ are a Markov chain with transition probabilities p_{ij} and initial probabilities π_i.

Suppose that the π_i (until now unspecified) are stationary: $\sum_i \pi_i p_{ij} = \pi_j$. Then

$$\sum_{u \in S} \pi_u p_{u u_1} p_{u_1 u_2} \cdots p_{u_{n-1} u_n} = \pi_{u_1} p_{u_1 u_2} \cdots p_{u_{n-1} u_n},$$

and it follows (see (24.4)) that T preserves P. Under the measure P specified by (24.6), T is the *Markov shift*.   ■

The shift T, qua point transformation on S^∞, is the same in Examples 24.1 and 24.2. A measure-preserving transformation, however, is the point transformation together with the σ-field with respect to which it is measurable and the measure it preserves.

Example 24.3. Let P be Lebesgue measure λ on the unit interval, and take $T\omega = 2\omega$ (mod 1):

$$T\omega = \begin{cases} 2\omega & \text{if } 0 < \omega \le \frac{1}{2}, \\ 2\omega - 1 & \text{if } \frac{1}{2} < \omega \le 1. \end{cases}$$

If ω has nonterminating dyadic expansion $\omega = .d_1(\omega)d_2(\omega)\ldots$, then $T\omega = .d_2(\omega)d_3(\omega)\ldots$: T shifts the digits one place to the left—compare (24.1). Since $T^{-1}(0, x] = (0, \frac{1}{2}x] \cup (\frac{1}{2}, \frac{1}{2} + \frac{1}{2}x]$, it follows by Lemma 1 that T preserves Lebesgue measure. This is the *dyadic* transformation.   ■

Example 24.4. Let Ω be the unit circle in the complex plane, let $\mathscr{F}$ be the σ-field generated by the arcs, and let P be normalized circular Lebesgue measure: map $[0, 1)$ to the unit circle by $\phi(x) = e^{2\pi i x}$ and define P by $P(A) = \lambda(\phi^{-1}A)$. For a fixed c in Ω, let $T\omega = c\omega$. Since T is effectively the *rotation* of the circle through the angle $\arg c$, T preserves P. The rotation turns out to have radically different properties according as c is a root of unity or not.   ■

Ergodicity

The $\mathscr{F}$-set A is *invariant* under T if $T^{-1}A = A$; it is a *nontrivial* invariant set if $0 < P(A) < 1$. And T is by definition *ergodic* if there are in $\mathscr{F}$ no

nontrivial invariant sets. A measurable function f is invariant if $f(T\omega) = f(\omega)$ for all ω; A is invariant if and only if I_A is.

The *ergodic theorem*:

Theorem 24.1. *Suppose that T is a measure-preserving transformation on $(\Omega, \mathscr{F}, P)$ and that f is measurable and integrable. Then*

$$(24.7) \qquad \lim_n \frac{1}{n} \sum_{k=1}^{n} f(T^{k-1}\omega) = \hat{f}(\omega)$$

with probability 1, where $\hat{f}$ is invariant and integrable and $E[\hat{f}] = E[f]$. If T is ergodic, then $\hat{f} = E[f]$ with probability 1.

This will be proved later in the section. In Section 34, $\hat{f}$ will be identified as a conditional expected value (see Example 34.3).

If $f = I_A$, (24.7) becomes

$$(24.8) \qquad \lim_n \frac{1}{n} \sum_{k=1}^{n} I_A(T^{k-1}\omega) = \hat{I}_A(\omega),$$

and in the ergodic case,

$$(24.9) \qquad \hat{I}_A(\omega) = P(A)$$

with probability 1. If A is invariant, then $\hat{I}_A(\omega)$ is 1 on A and 0 on A^c, and so the limit can certainly be nonconstant if T is not ergodic.

Example 24.5. Take $\Omega = \{a, b, c, d, e\}$ and $\mathscr{F} = 2^\Omega$. If T is the cyclic permutation $T = (a, b, c, d, e)$ and T preserves P, then P assigns equal probabilities to the five points. Since $\varnothing$ and Ω are the only invariant sets, T is ergodic. It is easy to check (24.8) and (24.9) directly.

The transformation $T = (a, b, c)(d, e)$, a product of two cycles, preserves P if and only if a, b, c have equal probabilities and d, e have equal probabilities. If the probabilities are all positive, then since $\{a, b, c\}$ is invariant, T is not ergodic. If, say, $A = \{a, d\}$, the limit in (24.8) is $\frac{1}{3}$ on $\{a, b, c\}$ and $\frac{1}{2}$ on $\{d, e\}$. This illustrates the essential role of ergodicity. ∎

The coordinate functions $z_n(\cdot)$ in Example 24.1 are independent, and hence by Kolmogorov's zero-one law every set in the tail field $\mathscr{T} = \bigcap_n \sigma(z_n, z_{n+1}, \ldots)$ has probability 0 or 1. (That the z_n take values in the abstract set S does not affect the arguments.) If $A \in \sigma(z_1, \ldots, z_k)$, then $T^{-n}A \in \sigma(z_{n+1}, \ldots, z_{n+k}) \subset \sigma(z_{n+1}, z_{n+2}, \ldots)$; since this is true for each k, $A \in \mathscr{F} = \sigma(z_1, z_2, \ldots)$ implies (Theorem 13.1(i)) $T^{-n}A \in \sigma(z_{n+1}, z_{n+2}, \ldots)$. For A invariant, it follows that $A \in \mathscr{T}$: The Bernoulli shift is ergodic. Thus

the ergodic theorem does imply that (24.5) holds with probability 1, whatever A may be.

The ergodicity of the Bernoulli shift can be proved in a different way. If $A = [(z_1, \ldots, z_n) = u]$ and $B = [(z_1, \ldots, z_k) = v]$ for an n-tuple u and a k-tuple v, and if P is given by (24.3), then $P(A \cap T^{-n}B) = P(A)P(B)$ because $T^{-n}B = [(z_{n+1}, \ldots, z_{n+k}) = v]$. Fix n and A, and use the π–λ theorem to show that this holds for all B in $\mathcal{F}$. If B is invariant, then $P(A \cap B) = P(A)P(B)$ holds for the A above, and hence (π–λ again) holds for all A. Taking $B = A$ shows that $P(B) = (P(B))^2$ for invariant B, and $P(B)$ is 0 or 1. This argument is very close to the proof of the zero-one law, but a modification of it gives a criterion for ergodicity that applies to the Markov shift and other transformations.

Lemma 2. *Suppose that $\mathcal{P} \subset \mathcal{F}_0 \subset \mathcal{F}$, where $\mathcal{F}_0$ is a field, every set in $\mathcal{F}_0$ is a finite or countable disjoint union of $\mathcal{P}$-sets, and $\mathcal{F}_0$ generates $\mathcal{F}$. Suppose further that there exists a positive c with this property: For each A in $\mathcal{P}$ there is an integer n_A such that*

$$(24.10) \qquad P(A \cap T^{-n_A}B) \geq cP(A)P(B)$$

for all B in $\mathcal{P}$. Then $T^{-1}C = C$ implies that $P(C)$ is 0 or 1.

It is convenient not to require that T preserve P; but if it does, then it is an ergodic measure-preserving transformation. In the argument just given, $\mathcal{P}$ consists of the thin cylinders, $n_A = n$ if $A = [(z_1, \ldots, z_n) = u]$, $c = 1$, and $\mathcal{F}_0 = \mathcal{C}_0$ is the class of cylinders.

PROOF. Since every $\mathcal{F}_0$-set is a disjoint union of $\mathcal{P}$-sets, (24.10) holds for $B \in \mathcal{F}_0$ (and $A \in \mathcal{P}$). Since for fixed A the class of B satisfying (24.10) is monotone, it contains $\mathcal{F}$ (Theorem 3.4). If B is invariant, it follows that $P(A \cap B) \geq cP(A)P(B)$ for A in $\mathcal{P}$. But then, by the same argument, the inequality holds for all A in $\mathcal{F}$. Take $A = B^c$: If B is invariant, then $P(B^c)P(B) = 0$ and hence $P(B)$ is 0 or 1. ∎

To treat the Markov shift, take $\mathcal{F}_0$ to consist of the cylinders and $\mathcal{P}$ to consist of the thin ones. If $A = [(z_1, \ldots, z_n) = (u_1, \ldots, u_n)]$, $n_A = n + m - 1$, and $B = [(z_1, \ldots, z_k) = (v_1, \ldots, v_k)]$, then

$$(24.11) \quad \begin{aligned} P(A)P(B) &= \pi_{u_1} p_{u_1 u_2} \cdots p_{u_{n-1} u_n} \pi_{i_1} p_{i_1 i_2} \cdots p_{i_{k-1} i_k}, \\ P(A \cap T^{-n_A}B) &= \pi_{u_1} p_{u_1 u_2} \cdots p_{u_{n-1} u_n} p^{(m)}_{u_n i_1} p_{i_1 i_2} \cdots p_{i_{k-1} i_k}. \end{aligned}$$

The lemma will apply if there exist an integer m and a positive c such that $p^{(m)}_{ij} \geq c\pi_j$ for all i and j. By Theorem 8.9 (or Lemma 2, p. 125), there is in the irreducible, aperiodic case an m such that all $p^{(m)}_{ij}$ are positive; take c

less than the minimum. By Lemma 2, the corresponding Markov shift is ergodic.

Example 24.6. Maps preserve ergodicity. Suppose that ψ: $\Omega \to \Omega$ is measurable $\mathcal{F}/\mathcal{F}$ and commutes with T in the sense that $\psi T\omega = T\psi\omega$. If T preserves P, it also preserves $P\psi^{-1}$: $P\psi^{-1}(T^{-1}A) = P(T^{-1}\psi^{-1}A) = P\psi^{-1}(A)$. And if T is ergodic under P, it is also ergodic under $P\psi^{-1}$: if A is invariant, so is $\psi^{-1}A$, and hence $P\psi^{-1}(A)$ is 0 or 1. These simple observations are useful in studying the ergodicity of stochastic processes (Theorem 36.4). ∎

Ergodicity of Rotations

The dyadic transformation, Example 24.3, is essentially the same as the Bernoulli shift. In any case, it is easy to use the zero-one law or Lemma 2 to show that it is ergodic. From this and the ergodic theorem, the normal number theorem follows once again.

Consider the rotations of Example 24.4. If the complex number c defining the rotation $(T\omega = c\omega)$ is -1, then the set consisting of the first and third quadrants is a nontrivial invariant set, and hence T is not ergodic. A similar construction shows that T is nonergodic whenever c is a root of unity.

In the opposite case, c is ergodic. In the first place, it is an old number-theoretic fact due to Kronecker that if c is not a root of unity then the orbit $(\omega, c\omega, c^2\omega, \dots)$ of every ω is dense. Since the orbits are rotations of one another, it suffices to prove that the orbit $(1, c, c^2, \dots)$ of 1 is dense. But if c is not a root of unity, then the elements of this orbit are all distinct and hence by compactness have a limit point ω_0. For arbitrary ϵ, there are distinct points c^n and c^{n+k} within $\epsilon/2$ of ω_0 and hence within ϵ of each other (distance measured along the arc). But then, since the distance from c^{n+jk} to $c^{n+(j+1)k}$ is the same as that from c^n to c^{n+k}, it is clear that for some m the points $c^n, c^{n+k}, \dots, c^{n+mk}$ form a chain which extends all the way around the circle and in which the distance from one point to the next is less than ϵ. Thus every point on the circle is within ϵ of some point of the orbit $(1, c, c^2, \dots)$, which is indeed dense.

To use this result to prove ergodicity, suppose that A is invariant and $P(A) > 0$. To show that $P(A)$ must then be 1, observe first that for arbitrary ϵ there is an arc I, of length at most ϵ, satisfying $P(A \cap I) > (1 - \epsilon)P(A)$. Indeed, A can be covered by a sequence $I_1, I_2, \dots$ of arcs for which $P(A)/(1 - \epsilon) > \sum_n P(I_n)$; the arcs can be taken disjoint and of length less than ϵ. Since $\sum_n P(A \cap I_n) = P(A) > (1 - \epsilon)\sum_n P(I_n)$, there is an n for which $P(A \cap I_n) > (1 - \epsilon)P(I_n)$: take $I = I_n$. Let I have length α; $\alpha \le \epsilon$.

Since A is invariant and T is invertible and preserves P, it follows that $P(A \cap T^nI) \ge (1 - \epsilon)P(T^nI)$. Suppose the arc I runs from a to b. Let n_1 be arbitrary and, using the fact that $\{T^na\}$ is dense, choose n_2 so that $T^{n_1}I$ and $T^{n_2}I$ are disjoint and the distance from $T^{n_1}b$ to $T^{n_2}a$ is less than $\epsilon\alpha$. Then choose n_3 so that $T^{n_1}I, T^{n_2}I, T^{n_3}I$ are disjoint and the distance from $T^{n_2}b$ to $T^{n_3}a$ is less than $\epsilon\alpha$. Continue until $T^{n_k}b$ is within α of $T^{n_1}a$ and a further step is impossible. Since the $T^{n_i}I$ are disjoint, $k\alpha \le 1$; and by the construction, the $T^{n_i}I$ cover the circle to within a set of measure $k\epsilon\alpha + \alpha$, which is at most 2ϵ. And now by disjointness,

$$P(A) \ge \sum_{i=1}^k P(A \cap T^{n_i}I) \ge (1 - \epsilon)\sum_{i=1}^k P(T^{n_i}I) \ge (1 - \epsilon)(1 - 2\epsilon).$$

Since ϵ was arbitrary, $P(A)$ must be 1: T is ergodic if c is not a root of unity.[†]

[†] For a simple Fourier-series proof, see Problem 26.30.

Proof of the Ergodic Theorem

The argument depends on a preliminary result the statement and proof of which are most clearly expressed in terms of functional operators. For a real function f on Ω, let Uf be the real function with value $(Uf)(\omega) = f(T\omega)$ at ω. If f is integrable, then by change of variable (Theorem 16.13),

$$(24.12) \quad E[Uf] = \int_\Omega f(T\omega) P(d\omega) = \int_\Omega f(\omega) PT^{-1}(d\omega) = E[f].$$

And the operator U is nonnegative in the sense that it carries nonnegative functions to nonnegative functions; hence $f \le g$ (pointwise) implies $Uf \le Ug$.

Make these pointwise definitions: $S_0 f = 0$, $S_n f = f + Uf + \cdots + U^{n-1} f$, $M_n f = \max_{0 \le k \le n} S_k f$, and $M_\infty f = \sup_{n \ge 0} S_n f = \sup_{n \ge 0} M_n f$. The *maximal ergodic theorem*:

Theorem 24.2. *If f is integrable, then*

$$(24.13) \qquad\qquad \int_{M_\infty f > 0} f \, dP \ge 0.$$

PROOF. Since $B_n = [M_n f > 0] \uparrow [M_\infty f > 0]$, it is enough, by the dominated convergence theorem, to show that $\int_{B_n} f \, dP \ge 0$. On B_n, $M_n f = \max_{1 \le k \le n} S_k f$. Since the operator U is nonnegative, $S_k f = f + US_{k-1} f \le f + UM_n f$ for $1 \le k \le n$, and therefore $M_n f \le f + UM_n f$ on B_n. This and the fact that the function $UM_n f$ is nonnegative imply

$$\int_\Omega M_n f \, dP = \int_{B_n} M_n f \, dP \le \int_{B_n} (f + UM_n f) \, dP$$

$$\le \int_{B_n} f \, dP + \int_\Omega UM_n f \, dP = \int_{B_n} f \, dP + \int_\Omega M_n f \, dP,$$

where the last equality follows from (24.12). Hence $\int_{B_n} f \, dP \ge 0$. ∎

Replace f by fI_A. If A is invariant, then $S_n(fI_A) = (S_n f)I_A$, and $M_\infty(fI_A) = (M_\infty f)I_A$, and therefore (24.13) gives

$$(24.14) \qquad \int_{A \cap [M_\infty f > 0]} f \, dP \ge 0 \qquad \text{if } T^{-1}A = A.$$

Now replace f here by $f - \lambda$, λ a constant. Clearly $[M_\infty(f - \lambda) > 0]$ is the set

where for some $n \geq 1$, $S_n(f - \lambda) > 0$. or $n^{-1}S_n f > \lambda$. Let

$$(24.15) \qquad F_\lambda = \left[\omega: \sup_{n \geq 1} \frac{1}{n} \sum_{k=1}^{n} f(T^{k-1}\omega) > \lambda \right];$$

it follows by (24.14) that $\int_{A \cap F_\lambda}(f - \lambda)\, dP \geq 0$, or

$$(24.16) \qquad \lambda P(A \cap F_\lambda) \leq \int_{A \cap F_\lambda} f\, dP \qquad \text{if } T^{-1}A = A.$$

The λ here can have either sign.

PROOF OF THEOREM 24.1. To prove that the averages $a_n(\omega) = n^{-1}\sum_{k=1}^{n} f(T^{k-1}\omega)$ converge, consider the set

$$A_{\alpha,\beta} = \left[\omega: \liminf_n a_n(\omega) < \alpha < \beta < \limsup_n a_n(\omega) \right]$$

for $\alpha < \beta$. Since $A_{\alpha,\beta} = A_{\alpha,\beta} \cap F_\beta$ and $A_{\alpha,\beta}$ is invariant, (24.16) gives $\beta P(A_{\alpha,\beta}) \leq \int_{A_{\alpha,\beta}} f\, dP$. The same result with $-f$, $-\beta$, $-\alpha$ in place of f, α, β is $\int_{A_{\alpha,\beta}} f\, dP \leq \alpha P(A_{\alpha,\beta})$. Since $\alpha < \beta$, the two inequalities together lead to $P(A_{\alpha,\beta}) = 0$. Take the union over rational α and β: The averages $a_n(\omega)$ converge, that is,

$$(24.17) \qquad \lim_n a_n(\omega) = \hat{f}(\omega)$$

with probability 1, where $\hat{f}$ may take the values $\pm\infty$ at certain values of ω.

Because of (24.12), $E[a_n] = E[f]$; if it is shown that the $a_n(\omega)$ are uniformly integrable, then it will follow (Theorem 16.14) that $\hat{f}$ is integrable and $E[\hat{f}] = E[f]$.

By (24.16), $\lambda P(F_\lambda) \leq E[|f|]$. Combine this with the same inequality for $-f$: If $G_\lambda = [\omega: \sup_n |a_n(\omega)| > \lambda]$, then $\lambda P(G_\lambda) \leq 2E[|f|]$ (trivial if $\lambda \leq 0$). Therefore, for positive α and λ,

$$\int_{[|a_n|>\lambda]} |a_n|\, dP \leq \frac{1}{n} \sum_{k=1}^{n} \int_{G_\lambda} |f(T^{k-1}\omega)|P(d\omega)$$

$$\leq \frac{1}{n} \sum_{k=1}^{n} \left(\int_{|f(T^{k-1}\omega)|>\alpha} |f(T^{k-1}\omega)|P(d\omega) + \alpha P(G_\lambda) \right)$$

$$= \int_{|f(\omega)|>\alpha} |f(\omega)|P(d\omega) + \alpha P(G_\lambda)$$

$$\leq \int_{|f(\omega)|>\alpha} |f(\omega)|P(d\omega) + 2\frac{\alpha}{\lambda}E[|f|].$$

Take $\alpha = \lambda^{1/2}$; since f is integrable, the final expression here goes to 0 as

$\lambda \to \infty$. The $a_n(\omega)$ are therefore uniformly integrable, and $E[\hat{f}] = E[f]$. The uniform integrability also implies $E[|a_n - \hat{f}|] \to 0$.

Set $\hat{f}(\omega) = 0$ outside the set where the $a_n(\omega)$ have a finite limit. Then (24.17) still holds with probability 1, and $\hat{f}(T\omega) = \hat{f}(\omega)$. Since $[\omega: \hat{f}(\omega) \le x]$ is invariant, in the ergodic case its measure is either 0 or 1; if x_0 is the infimum of the x for which it is 1, then $\hat{f}(\omega) = x_0$ with probability 1, and from $x_0 = E[\hat{f}] = E[f]$ it follows that $\hat{f}(\omega) = E[f]$ with probability 1. $\blacksquare$

The Continued-Fraction Transformation

Let Ω consist of the irrationals in the unit interval, and for x in Ω let $Tx = \{1/x\}$ and $a_1(x) = \lfloor 1/x \rfloor$ be the fractional and integral parts of $1/x$. This defines a mapping

$$(24.18) \qquad Tx = \left\{\frac{1}{x}\right\} = \frac{1}{x} - \left\lfloor\frac{1}{x}\right\rfloor = \frac{1}{x} - a_1(x)$$

of Ω into itself, a mapping associated with the continued-fraction expansion of x [A36]. Concentrating on *irrational* x avoids some trivial details connected with the rational case, where the expansion is finite; it is an inessential restriction because the interest here centers on results of the almost-everywhere kind.

For $x \in \Omega$ and $n \ge 1$ let $a_n(x) = a_1(T^{n-1}x)$ be the nth partial quotient, and define integer-valued functions $p_n(x)$ and $q_n(x)$ by the recursions

(24.19)

$$p_{-1}(x) = 1, \qquad p_0(x) = 0, \qquad p_n(x) = a_n(x)p_{n-1}(x) + p_{n-2}(x), \quad n \ge 1,$$
$$q_{-1}(x) = 0, \qquad q_0(x) = 1, \qquad q_n(x) = a_n(x)q_{n-1}(x) + q_{n-2}(x), \quad n \ge 1.$$

Simple induction arguments show that

$$(24.20) \qquad p_{n-1}(x)q_n(x) - p_n(x)q_{n-1}(x) = (-1)^n, \qquad n \ge 0,$$

and [A37: (27)]

$$(24.21) \quad x = 1\lfloor a_1(x) + \cdots + 1\lfloor a_{n-1}(x) + 1\lfloor a_n(x) + T^n x, \qquad n \ge 1.$$

It also follows inductively [A36: (26)] that

$$(24.22) \qquad 1\lfloor a_1(x) + \cdots + 1\lfloor a_{n-1}(x) + 1\lfloor a_n(x) + t$$
$$= \frac{p_n(x) + tp_{n-1}(x)}{q_n(x) + tq_{n-1}(x)}, \qquad n \ge 1, \quad 0 \le t \le 1.$$

Taking $t = 0$ here gives the formula for the nth convergent:

$$(24.23) \qquad \underline{1}\overline{|a_1(x)} + \cdots + \underline{1}\overline{|a_n(x)} = \frac{p_n(x)}{q_n(x)}, \qquad n \geq 1,$$

where, as follows from (24.20), $p_n(x)$ and $q_n(x)$ are relatively prime. By (24.21) and (24.22),

$$(24.24) \qquad x = \frac{p_n(x) + (T^n x)p_{n-1}(x)}{q_n(x) + (T^n x)q_{n-1}(x)}, \qquad n \geq 0,$$

which, together with (24.20), implies[†]

$$(24.25) \quad x - \frac{p_n(x)}{q_n(x)} = \frac{(-1)^n}{q_n(x)\left((T^n x)^{-1} q_n(x) + q_{n-1}(x)\right)}, \qquad n \geq 0.$$

Thus the convergents for even n fall to the left of x, and those for odd n fall to the right. And since (24.19) obviously implies that $q_n(x)$ goes to infinity with n, the convergents $p_n(x)/q_n(x)$ do converge to x: Each irrational x in $(0, 1)$ has the infinite simple continued-fraction representation

$$(24.26) \qquad x = \underline{1}\overline{|a_1(x)} + \underline{1}\overline{|a_2(x)} + \cdots .$$

The representation is unique [A36: (35)], and $Tx = \underline{1}\overline{|a_2(x)} + \underline{1}\overline{|a_3(x)} + \cdots$: T shifts the partial quotients in the same way the dyadic transformation (Example 24.3) shifts the digits of the dyadic expansion. Since the continued-fraction transformation turns out to be ergodic, it can be used to study the continued-fraction algorithm.

Suppose now that $a_1, a_2, \ldots$ are positive integers and define p_n and q_n by the recursions (24.19) without the argument x. Then (24.20) again holds (without the x), and so $p_n/q_n - p_{n-1}/q_{n-1} = (-1)^{n+1}/q_{n-1}q_n$, $n \geq 1$. Since q_n increases to infinity, the right side here is the nth term of a convergent alternating series. And since $p_0/q_0 = 0$, the nth partial sum is p_n/q_n, which therefore converges to some limit: Every simple infinite continued fraction converges, and [A36: (36)] the limit is an irrational in $(0, 1)$.

Let $\Delta_{a_1 \cdots a_n}$ be the set of x in Ω such that $a_k(x) = a_k$ for $1 \leq k \leq n$; call it a *fundamental set of rank n*. These sets are analogous to the dyadic intervals and the thin cylinders. For an explicit description of $\Delta_{a_1 \cdots a_n}$—necessary for the proof of ergodicity—consider the function

$$(24.27) \qquad \psi_{a_1 \ldots a_n}(t) = \underline{1}\overline{|a_1} + \cdots + \underline{1}\overline{|a_{n-1}} + \underline{1}\overline{|a_n + t}.$$

[†]Theorem 1.4 follows from this.

If $x \in \Delta_{a_1 \cdots a_n}$, then $x = \psi_{a_1 \cdots a_n}(T^n x)$ by (24.21); on the other hand, because of the uniqueness of the partial quotients [A36: (33)], if t is an irrational in the unit interval, then (24.27) lies in $\Delta_{a_1 \cdots a_n}$. Thus $\Delta_{a_1 \cdots a_n}$ is the image under (24.27) of Ω itself.

Just as (24.22) follows by induction, so does

$$(24.28) \qquad \psi_{a_1 \cdots a_n}(t) = \frac{p_n + t p_{n-1}}{q_n + t q_{n-1}}.$$

And $\psi_{a_1 \cdots a_n}(t)$ is increasing or decreasing in t according as n is even or odd, as is clear from the form of (24.27) (or differentiate in (24.28) and use (24.20)). It follows that

$$\Delta_{a_1 \cdots a_n} = \begin{cases} \left(\dfrac{p_n}{q_n}, \dfrac{p_n + p_{n-1}}{q_n + q_{n-1}} \right] \cap \Omega & \text{if } n \text{ is even,} \\[3ex] \left(\dfrac{p_n + p_{n-1}}{q_n + q_{n-1}}, \dfrac{p_n}{q_n} \right) \cap \Omega & \text{if } n \text{ is odd.} \end{cases}$$

By (24.20), this set has Lebesgue measure

$$(24.29) \qquad \lambda\left(\Delta_{a_1 \cdots a_n} \right) = \frac{1}{q_n(q_n + q_{n-1})}.$$

The fundamental sets of rank n form a partition of Ω, and their unions form a field $\mathcal{F}_n$; let $\mathcal{F}_0 = \bigcup_{n=1}^{\infty} \mathcal{F}_n$. Then $\mathcal{F}_0$ is the field generated by the class $\mathcal{P}$ of all the fundamental sets, and since each set in $\mathcal{F}_0$ is in some $\mathcal{F}_n$, each is a finite or countable disjoint union of $\mathcal{P}$-sets. Since $q_n \geq 2 q_{n-2}$ by (24.19), induction gives $q_n \geq 2^{(n-1)/2}$ for $n \geq 0$. And now (24.29) implies that $\mathcal{F}_0$ generates the σ-field $\mathcal{F}$ of linear Borel sets that are subsets of Ω (use Theorem 10.1(ii)). Thus $\mathcal{P}$, $\mathcal{F}_0$, $\mathcal{F}$ are related as in the hypothesis of Lemma 2. Clearly T is measurable $\mathcal{F}/\mathcal{F}$.

Although T does not preserve λ, it does preserve Gauss's measure, defined by

$$(24.30) \qquad P(A) = \frac{1}{\log 2} \int_A \frac{dx}{1+x}, \qquad A \in \mathcal{F}.$$

In fact, since

$$T^{-1}((0, t) \cap \Omega) = \bigcup_{k=1}^{\infty} \left(\left(\frac{1}{k+t}, \frac{1}{k} \right] \cap \Omega \right),$$

it is enough to verify

$$\int_0^1 \frac{dx}{1+x} = \sum_{k=1}^\infty \int_{1/(k+1)}^{1/k} \frac{dx}{1+x} = \sum_{k=1}^\infty \int_{1/(k+1)}^{1/k} \frac{dx}{1+x}.$$

Gauss's measure is useful because it is preserved by T and has the same sets of measure 0 as Lebesgue measure does.

Proof that T is ergodic. Fix $a_1,\ldots,a_n$, and write ψ_n for $\psi_{a_1\cdots a_n}$ and Δ_n for $\Delta_{a_1\cdots a_n}$. Suppose that n is even, so that ψ_n is increasing. If $x \in \Delta_n$, then (since $x = \psi_n(T^n x)$) $s < T^n x < t$ if and only if $\psi_n(s) < x < \psi_n(t)$; and this last condition implies $x \in \Delta_n$. Combined with (24.28) and (24.20), this shows that

$$\lambda(\Delta_n \cap [x: s < T^n x < t]) = \psi_n(t) - \psi_n(s) = \frac{t-s}{(q_n + sq_{n-1})(q_n + tq_{n-1})}.$$

If B is an interval with endpoints s and t, then by (24.29),

$$\lambda(\Delta_n \cap T^{-n}B) = \lambda(\Delta_n)\lambda(B)\frac{q_n(q_n + q_{n-1})}{(q_n + sq_{n-1})(q_n + tq_{n-1})}.$$

A similar argument establishes this for n odd. Since the ratio on the right lies between $\frac{1}{2}$ and 2,

(24.31) $\frac{1}{2}\lambda(\Delta_n)\lambda(B) \le \lambda(\Delta_n \cap T^{-n}B) \le 2\lambda(\Delta_n)\lambda(B).$

Therefore, (24.10) holds for $\mathscr{P}, \mathscr{F}_0, \mathscr{F}$ as defined above, $A = \Delta_n$, $n_A = n$, $c = \frac{1}{2}$, and λ in the role of P. Thus $T^{-1}C = C$ implies that $\lambda(C)$ is 0 or 1, and since Gauss's measure (24.30) comes from a density, $P(C)$ is 0 or 1 as well. Therefore, *T is an ergodic measure-preserving transformation on* $(\Omega, \mathscr{F}, P)$.

It follows by the ergodic theorem that if f is integrable, then

(24.32) $$\lim_n \frac{1}{n} \sum_{k=1}^n f(T^{k-1}x) = \frac{1}{\log 2} \int_0^1 \frac{f(x)}{1+x}\, dx$$

holds almost everywhere. Since the density in (24.30) is bounded away from 0 and ∞, the "integrable" and "almost everywhere" here can refer to P or to λ indifferently.

Taking f to be the indicator of the x-set where $a_1(x) = k$ shows that the asymptotic relative frequency of k among the partial quotients is almost everywhere equal to

$$\frac{1}{\log 2} \int_{1/(k+1)}^{1/k} \frac{dx}{1+x} = \frac{1}{\log 2} \log \frac{(k+1)^2}{k(k+2)}.$$

In particular, the partial quotients are unbounded almost everywhere.

For understanding the accuracy of the continued-fraction algorithm, the magnitude of $a_n(x)$ is less important than that of $q_n(x)$. The key relationship is

$$(24.33) \qquad \frac{1}{q_n(x)(q_n(x) + q_{n+1}(x))} < \left| x - \frac{p_n(x)}{q_n(x)} \right| < \frac{1}{q_n(x)q_{n+1}(x)},$$

which follows from (24.25) and (24.19). Suppose it is shown that

$$(24.34) \qquad \lim_n \frac{1}{n} \log q_n(x) = \frac{\pi^2}{12 \log 2}$$

almost everywhere; (24.33) will then imply that

$$(24.35) \qquad \lim_n \frac{1}{n} \log \left| x - \frac{p_n(x)}{q_n(x)} \right| = -\frac{\pi^2}{6 \log 2}:$$

The discrepancy between x and its nth convergent is almost everywhere of the order $e^{-n\pi^2/(6 \log 2)}$.

To prove (24.34), note first that since $p_{j+1}(x) = q_j(Tx)$ by (24.19), the product $\prod_{k=1}^{n} p_{n-k+1}(T^{k-1}x)/q_{n-k+1}(T^{k-1}x)$ telescopes to $1/q_n(x)$:

$$(24.36) \qquad \log \frac{1}{q_n(x)} = \sum_{k=1}^{n} \log \frac{p_{n-k+1}(T^{k-1}x)}{q_{n-k+1}(T^{k-1}x)}.$$

As observed earlier, $q_n(x) \geq 2^{(n-1)/2}$ for $n \geq 1$. Therefore, by (24.33),

$$\left| \frac{x}{p_n(x)/q_n(x)} - 1 \right| \leq \frac{1}{q_{n+1}(x)} \leq \frac{1}{2^{n/2}}, \quad n \geq 1.$$

Since $|\log(1 + s)| \leq 4|s|$ if $|s| \leq 1/\sqrt{2}$,

$$\left| \log x - \log \frac{p_n(x)}{q_n(x)} \right| \leq \frac{4}{2^{n/2}}.$$

Therefore, by (24.36),

$$\left| \sum_{k=1}^{n} \log T^{k-1}x - \log \frac{1}{q_n(x)} \right| \leq \sum_{k=1}^{n} \left| \log T^{k-1}x - \log \frac{p_{n-k+1}(T^{k-1}x)}{q_{n-k+1}(T^{k-1}x)} \right|$$

$$\leq \sum_{i=1}^{\infty} \frac{4}{2^{i/2}} < \infty.$$

By the ergodic theorem, then,[†]

$$\lim_n \frac{1}{n} \log \frac{1}{q_n(x)} = \frac{1}{\log 2} \int_0^1 \frac{\log x}{1+x} \, dx = \frac{-1}{\log 2} \int_0^1 \log(1+x) \, \frac{dx}{x}$$

$$= \frac{-1}{\log 2} \sum_{k=0}^\infty \frac{(-1)^k}{(k+1)^2} = \frac{-\pi^2}{12 \log 2}.$$

Hence (24.34).

Diophantine Approximation

The fundamental theorem of the measure theory of Diophantine approximation, due to Khinchine, is Theorem 1.5 together with Theorem 1.6. As in Section 1, let $\varphi(q)$ be a positive function of integers and let A_φ be the set of x in $(0,1)$ such that

(24.37) $$\left| x - \frac{p}{q} \right| < \frac{1}{q^2 \varphi(q)}$$

has infinitely many irreducible solutions p/q. If $\Sigma 1/q\varphi(q)$ converges, then A_φ has Lebesgue measure 0, as was proved in Section 1 (Theorem 1.6). It remains to prove that if φ is nondecreasing and $\Sigma 1/q\varphi(q)$ diverges, then A_φ has Lebesgue measure 1 (Theorem 1.5). It is enough to consider irrational x.

Lemma 3. *For positive α_n, the probability $(P$ or $\lambda)$ of $[x: a_n(x) > \alpha_n$ i.o.$]$ is 0 or 1 as $\Sigma 1/\alpha_n$ converges or diverges.*

PROOF.   Let $E_n = [x: a_n(x) > \alpha_n]$. Since $P(E_n) = P[x: a_1(x) > \alpha_n]$ is of the order $1/\alpha_n$, the first Borel–Cantelli lemma settles the convergent case (not needed in the proof of Theorem 1.5).
By (24.31),

$$\lambda(\Delta_n \cap E_{n+1}) \geq \tfrac{1}{2}\lambda(\Delta_n)\lambda[x: a_1(x) > \alpha_{n+1}] \geq \tfrac{1}{2}\lambda(\Delta_n)\frac{1}{\alpha_{n+1}+1}.$$

Taking a union over certain of the Δ_n shows that for $m < n$,

$$\lambda(E_m^c \cap \cdots \cap E_n^c \cap E_{n+1}) \geq \lambda(E_m^c \cap \cdots \cap E_n^c)\frac{1}{2(\alpha_{n+1}+1)}.$$

By induction on n,

$$\lambda(E_m^c \cap \cdots \cap E_n^c) \leq \prod_{k=m}^n \left(1 - \frac{1}{2(\alpha_{k+1}+1)}\right)$$

$$\leq \exp\left[-\sum_{k=m}^n \frac{1}{2(\alpha_{k+1}+1)}\right],$$

as in the proof of the second Borel–Cantelli lemma.   ∎

[†]Integrate by parts over $(\alpha, 1)$ and then let $\alpha \downarrow 0$. For the series, see Problem 26.28. The specific value of the limit in (24.34) is not needed for the application that follows.

PROOF OF THEOREM 1.5. Fix an integer N such that $\log N$ exceeds the limit in (24.34). Then, except on a set of measure 0,

$$(24.38) \qquad\qquad q_n(x) < N^n$$

holds for all but finitely many n. Since φ is nondecreasing,

$$\sum_{N^n \le q < N^{n+1}} \cdot \frac{1}{q\varphi(q)} \le \frac{N}{\varphi(N^n)},$$

and $\Sigma 1/\varphi(N^n)$ diverges if $\Sigma 1/q\varphi(q)$ does. By the lemma, outside a set of measure 0, $a_{n+1}(x) \ge \varphi(N^n)$ holds for infinitely many n. If this inequality and (24.38) both hold, then by (24.33) and the assumption that φ is nondecreasing,

$$\left| x - \frac{p_n(x)}{q_n(x)} \right| < \frac{1}{q_n(x)q_{n+1}(x)} \le \frac{1}{a_{n+1}(x)q_n^2(x)}$$

$$\le \frac{1}{\varphi(N^n)q_n^2(x)} \le \frac{1}{\varphi(q_n(x))q_n^2(x)}.$$

But $p_n(x)/q_n(x)$ is irreducible by (24.20).   ∎

PROBLEMS

24.1. Fix $(\Omega, \mathscr{F})$ and a T measurable $\mathscr{F}/\mathscr{F}$. The probability measures on $(\Omega, \mathscr{F})$ preserved by T form a convex set C. Show that T is ergodic under P if and only if P is an extreme point of C—cannot be represented as a proper convex combination of distinct elements of C.

24.2. Show that T is ergodic if and only if $n^{-1}\sum_{k=1}^{n-1}P(A \cap T^{-k}B) \to P(A)P(B)$ for all A and B (or all A and B in a π-system generating $\mathscr{F}$).

24.3. ↑ The transformation T is *mixing* if

$$(24.39) \qquad\qquad P(A \cap T^{-n}B) \to P(A)P(B)$$

for all A and B.

(a) Show that mixing implies ergodicity.

(b) Show that T is mixing if (24.39) holds for all A and B in a π-system generating $\mathscr{F}$.

(c) Show that the Bernoulli shift is mixing.

(d) Show that a cyclic permutation is ergodic but not mixing.

(e) Show that if c is not a root of unity, then the rotation (Example 24.4) is ergodic but not mixing.

24.4. ↑ Write $T^{-n}\mathscr{F} = [T^{-n}A: A \in \mathscr{F}]$, and call the σ-field $\mathscr{F}_\infty = \bigcap_{n=1}^\infty T^{-n}\mathscr{F}$ *trivial* if every set in it has probability either 0 or 1. (If T is invertible, $\mathscr{F}_\infty$ is $\mathscr{F}$ and hence is trivial only in uninteresting cases.)

(a) Show that if $\mathscr{F}_\infty$ is trivial, then T is ergodic. (A cyclic permutation is ergodic even though $\mathscr{F}_\infty$ is not trivial.)

(b) Show that if the hypotheses of Lemma 2 are satisfied, then $\mathscr{F}_\infty$ is trivial.

(c) It can be shown by martingale theory that if $\mathscr{F}_\infty$ is trivial, then T is mixing; see Problem 35.20. Reconsider Problem 24.3(c)

24.5. 8.35 24.4↑ (a) Show that the shift corresponding to an irreducible, aperiodic Markov chain is mixing. Do this first by Problem 8 35, then by Problem 24.4(b), (c).

(b) Show that if the chain is irreducible but has period greater than 1, then the shift is ergodic but not mixing.

(c) Suppose the state space splits into two closed, disjoint, nonempty subsets, and that the initial distribution (stationary) gives positive weight to each. Show that the corresponding shift is not ergodic.

24.6. Show that if T is ergodic and if f is nonnegative and $E[f] = \infty$, then $n^{-1}\sum_{k=1}^n f(T^{k-1}\omega) \to \infty$ with probability 1.

24.7. 24.3↑ Suppose that $P_0(A) = \int_A \delta \, dP$ for all A ($\delta \geq 0$) and that T is mixing with respect to P (T need not preserve P_0). Use (21.9) to prove

$$P_0(T^{-n}A) = \int_{T^{-n}A} \delta \, dP \to P(A).$$

24.8. 24.6↑ (a) Show that

$$\frac{1}{n} \sum_{k=1}^n a_k(x) \to \infty$$

and

$$\sqrt[n]{a_1(x) \cdots a_n(x)} \to \prod_{k=1}^\infty \left(1 + \frac{1}{k^2 + 2k}\right)^{(\log k)/(\log 2)}$$

almost everywhere.

(b) Show that

$$\frac{1}{n} \log\left(q_n(x)\left|x - \frac{p_n(x)}{q_n(x)}\right|\right) \to -\frac{\pi^2}{12\log 2}.$$

24.9. 24.4 24.7↑ (a) Show that the continued-fraction transformation is mixing.

(b) Show that

$$\lambda[x: T^n x \leq t] \to \frac{\log(1+t)}{\log 2}, \qquad 0 \leq t \leq 1.$$

CHAPTER 5

Convergence of Distributions

SECTION 25. WEAK CONVERGENCE

Many of the best-known theorems in probability have to do with the asymptotic behavior of distributions. This chapter covers both general methods for deriving such theorems and specific applications. The present section concerns the general limit theory for distributions on the real line, and the methods of proof use in an essential way the order structure of the line. For the theory in R^k, see Section 29.

Definitions

Distribution functions F_n were defined in Section 14 to *converge weakly* to the distribution function F if

$$(25.1) \qquad \lim_n F_n(x) = F(x)$$

for every continuity point x of F; this is expressed by writing $F_n \Rightarrow F$. Examples 14.1, 14.2, and 14.3 illustrate this concept in connection with the asymptotic distribution of maxima. Example 14.4 shows the point of allowing (25.1) to fail if F is discontinuous at x; see also Example 25.4. Theorem 25.8 and Example 25.9 show why this exemption is essential to a useful theory.

If μ_n and μ are the probability measures on $(R^1, \mathscr{R}^1)$ corresponding to F_n and F, then $F_n \Rightarrow F$ if and only if

$$(25.2) \qquad \lim_n \mu_n(A) = \mu(A)$$

for every A of the form $A = (-\infty, x]$ for which $\mu\{x\} = 0$—see (20.5). In this case the distributions themselves are said to converge weakly, which is expressed by writing $\mu_n \Rightarrow \mu$. Thus $F_n \Rightarrow F$ and $\mu_n \Rightarrow \mu$ are only different expressions of the same fact. From weak convergence it follows that (25.2) holds for many sets A besides half-infinite intervals; see Theorem 25.8.

Example 25.1. Let F_n be the distribution function corresponding to a unit mass at n: $F_n = I_{[n,\infty)}$. Then $\lim_n F_n(x) = 0$ for every x, so that (25.1) is satisfied if $F(x) \equiv 0$. But $F_n \Rightarrow F$ does not hold, because F is not a distribution function. Weak convergence is defined in this section only for functions F_n and F that rise from 0 at $-\infty$ to 1 at $+\infty$—that is, it is defined only for *probability* measures μ_n and μ.[†]   ∎

Example 25.2. *Poisson approximation to the binomial.* Let μ_n be the binomial distribution (20.6) for $p = \lambda/n$ and let μ be the Poisson distribution (20.7). For nonnegative integers k,

$$\mu_n\{k\} = \binom{n}{k}\left(\frac{\lambda}{n}\right)^k\left(1 - \frac{\lambda}{n}\right)^{n-k}$$

$$= \frac{\lambda^k(1 - \lambda/n)^n}{k!} \times \frac{1}{(1 - \lambda/n)^k} \prod_{i=0}^{k-1}\left(1 - \frac{i}{n}\right)$$

if $n \geq k$. As $n \to \infty$ the second factor on the right goes to 1 for fixed k, and so $\mu_n\{k\} \to \mu\{k\}$; this is a special case of Theorem 23.2. By the series form of Scheffé's theorem (the corollary to Theorem 16.12), (25.2) holds for every set A of nonnegative integers. Since the nonnegative integers support μ and the μ_n, (25.2) even holds for every linear Borel set A. Certainly μ_n converges weakly to μ in this case.   ∎

Example 25.3. Let μ_n correspond to a mass of n^{-1} at each point k/n, $k = 0, 1, \ldots, n - 1$; let μ be Lebesgue measure confined to the unit interval. The corresponding distribution functions satisfy $F_n(x) = (\lfloor nx \rfloor + 1)/n \to F(x)$ for $0 \leq x < 1$, and so $F_n \Rightarrow F$. In this case (25.1) holds for every x, but (25.2) does not, as in the preceding example, hold for every Borel set A: if A is the set of rationals, then $\mu_n(A) = 1$ does not converge to $\mu(A) = 0$. Despite this, μ_n does converge weakly to μ.   ∎

Example 25.4. If μ_n is a unit mass at x_n and μ is a unit mass at x, then $\mu_n \Rightarrow \mu$ if and only if $x_n \to x$. If $x_n > x$ for infinitely many n, then (25.1) fails at the discontinuity point of F.   ∎

Uniform Distribution Modulo 1[*]

For a sequence $x_1, x_2, \ldots$ of real numbers, consider the corresponding sequence of their fractional parts $\{x_n\} = x_n - \lfloor x_n \rfloor$. For each n, define a probability measure μ_n by

$$(25.3) \qquad \mu_n(A) = \frac{1}{n}\#[k: 1 \leq k \leq n, \{x_k\} \in A];$$

[†]There is (see Section 28) a related notion of *vague* convergence in which μ may be *defective* in the sense that $\mu(R^1) < 1$. Weak convergence is in this context sometimes called *complete convergence*.

[*]This topic, which requires ergodic theory, may be omitted.

μ_n has mass n^{-1} at the points $\{x_1\}, \ldots, \{x_n\}$, and if several of these points coincide, the masses add. The problem is to find the weak limit of $\{\mu_n\}$ in number-theoretically interesting cases.

If the μ_n defined by (25.3) converge weakly to Lebesgue measure restricted to the unit interval, the sequence $x_1, x_2, \ldots$ is said to be *uniformly distributed modulo* 1. In this case every subinterval has asymptotically its proportional share of the points $\{x_n\}$; by Theorem 25.8 below, the same is then true of every subset whose boundary has Lebesgue measure 0.

Theorem 25.1. *For θ irrational, $\theta, 2\theta, 3\theta, \ldots$ is uniformly distributed modulo* 1.

PROOF.   Since $\{n\theta\} = \{n\{\theta\}\}$, θ can be assumed to lie in $[0,1]$. As in Example 24.4, map $[0, 1)$ to the unit circle in the complex plane by $\phi(x) = e^{2\pi i x}$. If θ is irrational, then $c = \phi(\theta)$ is not a root of unity, and so (p. 000) $T\omega = c\omega$ defines an ergodic transformation with respect to circular Lebesgue measure P. Let $\mathscr{J}$ be the class of open arcs with endpoints in some fixed countable, dense set. By the ergodic theorem, the orbit $\{T^n\omega\}$ of almost every ω enters every I in $\mathscr{J}$ with asymptotic relative frequency $P(I)$. Fix such an ω. If $I_1 \subset J \subset I_2$, where J is a closed arc and I_1, I_2 are in $\mathscr{J}$, then the upper and lower limits of $n^{-1}\sum_{k=1}^{n} I_J(T^{k-1}\omega)$ are between $P(I_1)$ and $P(I_2)$, and therefore the limit exists and equals $P(J)$. Since the orbits and the arcs are rotations of one another, *every* orbit enters *every* closed arc J with frequency $P(J)$. This is true in particular of the orbit $\{c^n\}$ of 1.

Now carry all this back to $[0, 1)$ by ϕ^{-1}: For every x in $[0, 1)$, $\{n\theta\} = \phi^{-1}(c^n)$ lies in $[0, x]$ with asymptotic relative frequency x.   ∎

For a simple proof by Fourier series, see Example 26.3.

Convergence in Distribution

Let X_n and X be random variables with respective distribution functions F_n and F. If $F_n \Rightarrow F$, then X_n is said to *converge in distribution* or *in law* to X, written $X_n \Rightarrow X$. This dual use of the double arrow will cause no confusion. Because of the defining conditions (25.1) and (25.2), $X_n \Rightarrow X$ if and only if

$$(25.4) \qquad\qquad \lim_n P[X_n \le x] = P[X \le x]$$

for every x such that $P[X = x] = 0$.

Example 25.5.   Let $X_1, X_2, \ldots$ be independent random variables, each with the exponential distribution: $P[X_n \ge x] = e^{-\alpha x}$, $x \ge 0$. Put $M_n = \max\{X_1, \ldots, X_n\}$ and $b_n = \alpha^{-1} \log n$. The relation (14.9), established in Example 14.1, can be restated as $P[M_n - b_n \le x] \to e^{-e^{-\alpha x}}$. If X is any random variable with distribution function $e^{-e^{-\alpha x}}$, this can be written $M_n - b_n \Rightarrow X$.
■

One is usually interested in proving weak convergence of the distributions of some given sequence of random variables, such as the $M_n - b_n$ in this example, and the result is often most clearly expressed in terms of the random variables themselves rather than in terms of their distributions or

distribution functions. Although the $M_n - b_n$ here arise naturally from the problem at hand, the random variable X is simply constructed to make it possible to express the asymptotic relation compactly by $M_n - b_n \Rightarrow X$. Recall that by Theorem 14.1 there does exist a random variable for any prescribed distribution.

Example 25.6. For each n, let Ω_n be the space of n-tuples of 0's and 1's, let $\mathcal{F}_n$ consist of all subsets of Ω_n, and let P_n assign probability $(\lambda/n)^k(1 - \lambda/n)^{n-k}$ to each ω consisting of k 1's and $n - k$ 0's. Let $X_n(\omega)$ be the number of 1's in ω; then X_n, a random variable on $(\Omega_n, \mathcal{F}_n, P_n)$, represents the number of successes in n Bernoulli trials having probability λ/n of success at each.

Let X be a random variable, on some $(\Omega, \mathcal{F}, P)$, having the Poisson distribution with parameter λ. According to Example 25.2, $X_n \Rightarrow X$. ∎

As this example shows, the random variables X_n may be defined on entirely different probability spaces. To allow for this possibility, the P on the left in (25.4) really should be written P_n. Suppressing the n causes no confusion if it is understood that P refers to whatever probability space it is that X_n is defined on; the underlying probability space enters into the definition only via the distribution μ_n it induces on the line. Any instance of $F_n \Rightarrow F$ or of $\mu_n \Rightarrow \mu$ can be rewritten in terms of convergence in distribution: There exist random variables X_n and X (on some probability spaces) with distribution functions F_n and F, and $F_n \Rightarrow F$ and $X_n \Rightarrow X$ express the same fact.

Convergence in Probability

Suppose that $X, X_1, X_2, \ldots$ are random variables all defined on the same probability space $(\Omega, \mathcal{F}, P)$. If $X_n \to X$ with probability 1, then $P[|X_n - X| \geq \epsilon \text{ i.o.}] = 0$ for $\epsilon > 0$, and hence

$$(25.5) \qquad \lim_n P\{|X_n - X| > \epsilon\} = 0$$

by Theorem 4.1. Thus there is *convergence in probability* $X_n \to_P X$; see Theorems 5.2 and 20.5.

Suppose that (25.5) holds for each positive ϵ. Now $P[X \leq x - \epsilon] - P[|X_n - X| \geq \epsilon] \leq P[X_n \leq x] \leq P[X \leq x + \epsilon] + P[|X_n - x| \geq \epsilon]$; letting n tend to ∞ and then letting ϵ tend to 0 shows that $P[X < x] \leq \liminf_n P[X_n \leq x] \leq \limsup_n P[X_n \leq x] \leq P[X \leq x]$. Thus $P[X_n \leq x] \to P[X \leq x]$ if $P[X = x] = 0$, and so $X_n \Rightarrow X$:

Theorem 25.2. *Suppose that X_n and X are random variables on the same probability space. If $X_n \to X$ with probability 1, then $X_n \to_P X$. If $X_n \to_P X$, then $X_n \Rightarrow X$.*

Of the two implications in this theorem, neither converse holds. Because of Example 5.4, convergence in probability does not imply convergence with probability 1. Neither does convergence in distribution imply convergence in probability: if X and Y are independent and assume the values 0 and 1 with probability $\frac{1}{2}$ each, and if $X_n = Y$, then $X_n \Rightarrow X$, but $X_n \to_p X$ cannot hold because $P[|X - Y|] = 1 = \frac{1}{2}$. What is more, (25.5) is impossible if X and the X_n are defined on different probability spaces, as may happen in the case of convergence in distribution.

Although (25.5) in general makes no sense unless X and the X_n are defined on the same probability space, suppose that X is replaced by a constant real number a—that is, suppose that $X(\omega) \equiv a$. Then (25.5) becomes

$$(25.6) \qquad\qquad \lim_n P[|X_n - a| \geq \epsilon] = 0,$$

and this condition makes sense even if the space of X_n does vary with n. Now a can be regarded as a random variable (on any probability space at all), and it is easy to show that (25.6) implies that $X_n \Rightarrow a$: Put $\epsilon = |x - a|$; if $x > a$, then $P[X_n \leq x] \geq P[|X_n - a| < \epsilon] \to 1$, and if $x < a$, then $P[X_n \leq x] \leq P[|X_n - a| \geq \epsilon] \to 0$. If a is regarded as a random variable, its distribution function is 0 for $x < a$ and 1 for $x \geq a$. Thus (25.6) implies that the distribution function of X_n converges weakly to that of a.

Suppose, on the other hand, that $X_n \Rightarrow a$. Then $P[|X_n - a| > \epsilon] \leq P[X_n \leq a - \epsilon] + 1 - P[X_n \leq a + \epsilon] \to 0$, so that (25.6) holds:

Theorem 25.3. *The condition* (25.6) *holds for all positive ϵ if and only if $X_n \Rightarrow a$, that is, if and only if*

$$\lim_n P[X_n \leq x] = \begin{cases} 0 & \text{if } x < a, \\ 1 & \text{if } x > a. \end{cases}$$

If (25.6) holds for all positive ϵ, X_n may be said to *converge to a in probability*. As this does not require that the X_n be defined on the same space, it is not really a special case of convergence in probability as defined by (25.5). Convergence in probability in this new sense will be denoted $X_n \Rightarrow a$, in accordance with the theorem just proved.

Example 14.4 restates the weak law of large numbers in terms of this concept. Indeed, if $X_1, X_2, \ldots$ are independent, identically distributed random variables with finite mean m, and if $S_n = X_1 + \cdots + X_n$, the weak law of large numbers is the assertion $n^{-1}S_n \Rightarrow m$. Example 6.3 provides another illustration: If S_n is the number of cycles in a random permutation on n letters, then $S_n/\log n \Rightarrow 1$.

Example 25.7. Suppose that $X_n \Rightarrow X$ and $\delta_n \to 0$. Given ϵ and η, choose x so that $P[|X| \geq x] < \eta$ and $P[X = \pm x] = 0$, and then choose n_0 so that $n \geq n_0$ implies that $|\delta_n| < \epsilon/x$ and $|P[X_n \leq y] - P[X \leq y]| < \eta$ for $y = \pm x$. Then $P[|\delta_n X_n| \geq \epsilon] < 3\eta$ for $n \geq n_0$. Thus $X_n \Rightarrow X$ *and* $\delta_n \to 0$ *imply that* $\delta_n X_n \Rightarrow 0$, a restatement of Lemma 2 of Section 14 (p. 193).   ∎

The asymptotic properties of a random variable should remain unaffected if it is altered by the addition of a random variable that goes to 0 in probability. Let (X_n, Y_n) be a two-dimensional random vector.

Theorem 25.4. *If* $X_n \Rightarrow X$ *and* $X_n - Y_n \Rightarrow 0$, *then* $Y_n \Rightarrow X$.

PROOF. Suppose that $y' < x < y''$ and $P[X = y'] = P[X = y''] = 0$. If $y' < x - \epsilon < x < x + \epsilon < y''$, then

$$(25.7) \quad P[X_n \leq y'] - P[|X_n - Y_n| \geq \epsilon] \leq P[Y_n \leq x]$$

$$\leq P[X_n \leq y''] + P[|X_n - Y_n| \geq \epsilon].$$

Since $X_n \Rightarrow X$, letting $n \to \infty$ gives

$$(25.8) \qquad P[X \leq y'] \leq \lim_n \inf P[Y_n \leq x]$$

$$\leq \lim_n \sup P[Y_n \leq x] \leq P[X \leq y''].$$

Since $P[X = y] = 0$ for all but countably many y, if $P[X = x] = 0$, then y' and y'' can further be chosen so that $P[X \leq y']$ and $P[X \leq y'']$ are arbitrarily near $P[X \leq x]$; hence $P[Y_n \leq x] \to P[X \leq x]$.   ∎

Theorem 25.4 has a useful extension. Suppose that $(X_n^{(u)}, Y_n)$ is a two-dimensional random vector.

Theorem 25.5. *If, for each* u, $X_n^{(u)} \Rightarrow X^{(u)}$ *as* $n \to \infty$, *if* $X^{(u)} \Rightarrow X$ *as* $u \to \infty$, *and if*

$$(25.9) \qquad \lim_u \lim_n \sup P[|X_n^{(u)} - Y_n| \geq \epsilon] = 0$$

for positive ϵ, *then* $Y_n \Rightarrow X$.

PROOF. Replace X_n by $X_n^{(u)}$ in (25.7). If $P[X = y'] = 0 \equiv P[X^{(u)} = y']$ and $P[X = y''] = 0 \equiv P[X^{(u)} = y'']$, letting $n \to \infty$ and then $u \to \infty$ gives (25.8) once again. Since $P[X = y] = 0 \equiv P[X^{(u)} = y]$ for all but countably many y, the proof can be completed as before.   ∎

Fundamental Theorems

Some of the fundamental properties of weak convergence were established in Section 14. It was shown there that a sequence cannot have two distinct weak limits: *If $F_n \Rightarrow F$ and $F_n \Rightarrow G$, then $F = G$.* The proof is simple: The hypothesis implies that F and G agree at their common points of continuity, hence at all but countably many points, and hence by right continuity at all points. Another simple fact is this: *If $\lim_n F_n(d) = F(d)$ for d in a set D dense in R^1, then $F_n \Rightarrow F$.* Indeed, if F is continuous at x, there are in D points d' and d'' such that $d' < x < d''$ and $F(d'') - F(d') < \epsilon$, and it follows that the limits superior and inferior of $F_n(x)$ are within ϵ of $F(x)$.

For any probability measure on $(R^1, \mathscr{R}^1)$ there is on some probability space a random variable having that measure as its distribution. Therefore, for probability measures satisfying $\mu_n \Rightarrow \mu$, there exist random variables Y_n and Y having these measures as distributions and satisfying $Y_n \Rightarrow Y$. According to the following theorem, the Y_n and Y can be constructed on the same probability space, and even in such a way that $Y_n(\omega) \to Y(\omega)$ for every ω—a condition much stronger than $Y_n \Rightarrow Y$. This result, *Skorohod's theorem*, makes possible very simple and transparent proofs of many important facts.

Theorem 25.6. *Suppose that μ_n and μ are probability measures on $(R^1, \mathscr{R}^1)$ and $\mu_n \Rightarrow \mu$. There exist random variables Y_n and Y on a common probability space $(\Omega, \mathscr{F}, P)$ such that Y_n has distribution μ_n, Y has distribution μ, and $Y_n(\omega) \to Y(\omega)$ for each ω.*

PROOF. For the probability space $(\Omega, \mathscr{F}, P)$, take $\Omega = (0, 1)$, let $\mathscr{F}$ consist of the Borel subsets of $(0, 1)$, and for $P(A)$ take the Lebesgue measure of A.

The construction is related to that in the proofs of Theorems 14.1 and 20.4. Consider the distribution functions F_n and F corresponding to μ_n and μ. For $0 < \omega < 1$, put $Y_n(\omega) = \inf[x: \omega \le F_n(x)]$ and $Y(\omega) = \inf[x: \omega \le F(x)]$. Since $\omega \le F_n(x)$ if and only if $Y_n(\omega) \le x$ (see the argument following (14.5)), $P[\omega: Y_n(\omega) \le x] = P[\omega: \omega \le F_n(x)] = F_n(x)$. Thus Y_n has distribution function F_n; similarly, Y has distribution function F.

It remains to show that $Y_n(\omega) \to Y(\omega)$. The idea is that Y_n and Y are essentially inverse functions to F_n and F; if the direct functions converge, so must the inverses.

Suppose that $0 < \omega < 1$. Given ϵ, choose x so that $Y(\omega) - \epsilon < x < Y(\omega)$ and $\mu\{x\} = 0$. Then $F(x) < \omega$; $F_n(x) \to F(x)$ now implies that, for n large enough, $F_n(x) < \omega$ and hence $Y(\omega) - \epsilon < x < Y_n(\omega)$. Thus $\liminf_n Y_n(\omega) \ge Y(\omega)$. If $\omega < \omega'$ and ϵ is positive, choose a y for which $Y(\omega') < y < Y(\omega') + \epsilon$ and $\mu\{y\} = 0$. Now $\omega < \omega' \le F(Y(\omega')) \le F(y)$, and so, for n large enough, $\omega \le F_n(y)$ and hence $Y_n(\omega) \le y < Y(\omega') + \epsilon$. Thus $\limsup_n Y_n(\omega) \le Y(\omega')$ if $\omega < \omega'$. Therefore, $Y_n(\omega) \to Y(\omega)$ if Y is continuous at ω.

Since Y is nondecreasing on $(0,1)$, it has at most countably many discontinuities. At discontinuity points ω of Y, redefine $Y_n(\omega) = Y(\omega) = 0$. With this change, $Y_n(\omega) \to Y(\omega)$ for every ω. Since Y and the Y_n have been altered only on a set of Lebesgue measure 0, their distributions are still μ_n and μ. ∎

Note that this proof uses the order structure of the real line in an essential way. The proof of the corresponding result in R^k is more complicated.

The following *mapping theorem* is of very frequent use.

Theorem 25.7. *Suppose that h: $R^1 \to R^1$ is measurable and that the set D_h of its discontinuities is measurable.† If $\mu_n \Rightarrow \mu$ and $\mu(D_h) = 0$, then $\mu_n h^{-1} \Rightarrow \mu h^{-1}$.*

Recall (see (13.7)) that μh^{-1} has value $\mu(h^{-1}A)$ at A.

PROOF. Consider the random variables Y_n and Y of Theorem 25.6. Since $Y_n(\omega) \to Y(\omega)$, if $Y(\omega) \notin D_h$ then $h(Y_n(\omega)) \to h(Y(\omega))$. Since $P[\omega\colon Y(\omega) \in D_h] = \mu(D_h) = 0$, it follows that $h(Y_n(\omega)) \to h(Y(\omega))$ with probability 1. Hence $h(Y_n) \Rightarrow h(Y)$ by Theorem 25.2. Since $P[h(Y) \in A] = P[Y \in h^{-1}A] = \mu(h^{-1}A)$, $h(Y)$ has distribution μh^{-1}; similarly, $h(Y_n)$ has distribution $\mu_n h^{-1}$. Thus $h(Y_n) \Rightarrow h(Y)$ is the same thing as $\mu_n h^{-1} \Rightarrow \mu h^{-1}$. ∎

Because of the definition of convergence in distribution, this result has an equivalent statement in terms of random variables:

Corollary 1. *If $X_n \Rightarrow X$ and $P[X \in D_h] = 0$, then $h(X_n) \Rightarrow h(X)$.*

Take $X \equiv a$:

Corollary 2. *If $X_n \Rightarrow a$ and h is continuous at a, then $h(X_n) \Rightarrow h(a)$.*

Example 25.8. From $X_n \Rightarrow X$ it follows directly by the theorem that $aX_n + b \Rightarrow aX + b$. Suppose also that $a_n \to a$ and $b_n \to b$. Then $(a_n - a)X_n \Rightarrow 0$ by Example 25.7, and so $(a_n X_n + b_n) - (aX_n + b) \Rightarrow 0$. And now $a_n X_n + b_n \Rightarrow aX + b$ follows by Theorem 25.4: If $X_n \Rightarrow X$, $a_n \to a$, and $b_n \to b$, then $a_n X_n + b_n \Rightarrow aX + b$. This fact was stated and proved differently in Section 14 —see Lemma 1 on p. 193. ∎

By definition, $\mu_n \Rightarrow \mu$ means that the corresponding distribution functions converge weakly. The following theorem characterizes weak convergence

---

†That D_h lies in $\mathscr{R}^1$ is generally obvious in applications. In point of fact, it always holds (even if h is not measurable): Let $A(\epsilon, \delta)$ be the set of x for which there exist y and z such that $|x - y| < \delta$, $|x - z| < \delta$, and $|h(y) - h(z)| \geq \epsilon$. Then $A(\epsilon, \delta)$ is open and $D_h = \bigcup_\epsilon \bigcap_\delta A(\epsilon, \delta)$, where ϵ and δ range over the positive rationals.

without reference to distribution functions. The boundary ∂A of A consists of the points that are limits of sequences in A and are also limits of sequences in A^c; alternatively, ∂A is the closure of A minus its interior. A set A is a μ-continuity set if it is a Borel set and $\mu(\partial A) = 0$.

Theorem 25.8. *The following three conditions are equivalent.*

(i) $\mu_n \Rightarrow \mu$;

(ii) $\int f d\mu_n \to \int f d\mu$ *for every bounded, continuous real function f*;

(iii) $\mu_n(A) \to \mu(A)$ *for every μ-continuity set A.*

PROOF. Suppose that $\mu_n \Rightarrow \mu$. and consider the random variables Y_n and Y of Theorem 25.6. Suppose that f is a bounded function such that $\mu(D_f) = 0$, where D_f is the set of points of discontinuity of F. From $P[Y \in D_f] = \mu(D_f) = 0$ it follows that $f(Y_n) \to f(Y)$ with probability 1, and so by change of variable (see (21.1)) and the bounded convergence theorem, $\int f d\mu_n = E[f(Y_n)] \to E[f(Y)] = \int f d\mu$. Thus $\mu_n \Rightarrow \mu$ and $\mu(D_f) = 0$ together imply that $\int f d\mu_n \to \int f d\mu$ if f is bounded. In particular, (i) implies (ii). Further, if $f = I_A$, then $D_f = \partial A$, and from $\mu(\partial A) = 0$ and $\mu_n \Rightarrow \mu$ follows $\mu_n(A) = \int f d\mu_n \to \int f d\mu = \mu(A)$. Thus (i) also implies (iii).

Since $\partial(-\infty, x] = \{x\}$, obviously (iii) implies (i). It therefore remains only to deduce $\mu_n \Rightarrow \mu$ from (ii). Consider the corresponding distribution functions. Suppose that $x < y$, and let $f(t)$ be 1 for $t \le x$, 0 for $t \ge y$, and interpolate linearly on $[x, y]$: $f(t) = (y - t)/(y - x)$ for $x \le t \le y$. Since $F_n(x) \le \int f d\mu_n$ and $\int f d\mu \le F(y)$, it follows from (ii) that $\limsup_n F_n(x) \le F(y)$; letting $y \downarrow x$ shows that $\limsup_n F_n(x) \le F(x)$. Similarly, $F(u) \le \liminf_n F_n(x)$ for $u < x$ and hence $F(x-) \le \liminf_n F_n(x)$. This implies convergence at continuity points. ∎

The function f in this last part of the proof is uniformly continuous. Hence $\mu_n \Rightarrow \mu$ follows if $\int f d\mu_n \to \int f d\mu$ for every bounded and *uniformly* continuous f.

Example 25.9. The distributions in Example 25.3 satisfy $\mu_n \Rightarrow \mu$, but $\mu_n(A)$ does not converge to $\mu(A)$ if A is the set of rationals. Hence this A cannot be a μ-continuity set; in fact, of course, $\partial A = R^1$. ∎

The concept of weak convergence would be nearly useless if (25.2) were not allowed to fail when $\mu(\partial A) > 0$. Since $F(x) - F(x-) = \mu\{x\} = \mu(\partial(-\infty, x])$, it is therefore natural in the original definition to allow (25.1) to fail when x is not a continuity point of F.

Helly's Theorem

One of the most frequently used results in analysis is the *Helly selection theorem*:

Theorem 25.9. *For every sequence* $\{F_n\}$ *of distribution functions there exists a subsequence* $\{F_{n_k}\}$ *and a nondecreasing, right-continuous function* F *such that* $\lim_k F_{n_k}(x) = F(x)$ *at continuity points* x *of* F.

PROOF. An application of the diagonal method [A14] gives a sequence $\{n_k\}$ of integers along which the limit $G(r) = \lim_k F_{n_k}(r)$ exists for every rational r. Define $F(x) = \inf[G(r): x < r]$. Clearly F is nondecreasing.

To each x and ϵ there is an r for which $x < r$ and $G(r) < F(x) + \epsilon$. If $x \le y < r$, then $F(y) \le G(r) < F(x) + \epsilon$. Hence F is continuous from the right.

If F is continuous at x, choose $y < x$ so that $F(x) - \epsilon < F(y)$; now choose rational r and s so that $y < r < x < s$ and $G(s) < F(x) + \epsilon$. From $F(x) - \epsilon < G(r) \le G(s) < F(x) + \epsilon$ and $F_n(r) \le F_n(x) \le F_n(s)$ it follows that as k goes to infinity $F_{n_k}(x)$ has limits superior and inferior within ϵ of $F(x)$.   ∎

The F in this theorem necessarily satisfies $0 \le F(x) \le 1$. But F need not be a distribution function: if F_n has a unit jump at n, for example, $F(x) \equiv 0$ is the only possibility. It is important to have a condition which ensures that for some subsequence the limit F is a distribution function.

A sequence of probability measures μ_n on $(R^1, \mathscr{R}^1)$ is said to be *tight* if for each ϵ there exists a finite interval $(a, b]$ such that $\mu_n(a, b] > 1 - \epsilon$ for all n. In terms of the corresponding distribution functions F_n, the condition is that for each ϵ there exist x and y such that $F_n(x) < \epsilon$ and $F_n(y) > 1 - \epsilon$ for all n. If μ_n is a unit mass at n, $\{\mu_n\}$ is not tight in this sense—the mass of μ_n "escapes to infinity." Tightness is a condition preventing this escape of mass.

Theorem 25.10. *Tightness is a necessary and sufficient condition that for every subsequence* $\{\mu_{n_k}\}$ *there exist a further subsequence* $\{\mu_{n_{k(j)}}\}$ *and a probability measure* μ *such that* $\mu_{n_{k(j)}} \Rightarrow \mu$ *as* $j \to \infty$.

Only the sufficiency of the condition in this theorem is used in what follows.

PROOF. *Sufficiency.* Apply Helly's theorem to the subsequence $\{F_{n_k}\}$ of corresponding distribution functions. There exists a further subsequence $\{F_{n_{k(j)}}\}$ such that $\lim_j F_{n_{k(j)}}(x) = F(x)$ at continuity points of F, where F is nondecreasing and right-continuous. There exists by Theorem 12.4 a measure μ on $(R^1, \mathscr{R}^1)$ such that $\mu(a, b] = F(b) - F(a)$. Given ϵ, choose a and b so that $\mu_n(a, b] > 1 - \epsilon$ for all n, which is possible by tightness. By decreasing a

and increasing b, one can ensure that they are continuity points of F. But then $\mu(a,b] \geq 1 - \epsilon$. Therefore, μ is a *probability* measure, and of course $\mu_{n_{k(j)}} \Rightarrow \mu$.

Necessity. If $\{\mu_n\}$ is not tight, there exists a positive ϵ such that for each finite interval $(a,b]$, $\mu_n(a,b] \leq 1 - \epsilon$ for some n. Choose n_k so that $\mu_{n_k}(-k,k] \leq 1 - \epsilon$. Suppose that some subsequence $\{\mu_{n_{k(j)}}\}$ of $\{\mu_{n_k}\}$ were to converge weakly to some probability measure μ. Choose $(a,b]$ so that $\mu\{a\} = \mu\{b\} = 0$ and $\mu(a, b] > 1 - \epsilon$. For large enough j, $(a, b] \subset (-k(j), k(j)]$, and so $1 - \epsilon \geq \mu_{n_{k(j)}}(-k(j), k(j)] \geq \mu_{n_{k(j)}}(a, b] \to \mu(a, b]$. Thus $\mu(a, b] \leq 1 - \epsilon$, a contradiction. ∎

Corollary. *If $\{\mu_n\}$ is a tight sequence of probability measures, and if each subsequence that converges weakly at all converges weakly to the probability measure μ, then $\mu_n \Rightarrow \mu$.*

PROOF. By the theorem, each subsequence $\{\mu_{n_k}\}$ contains a further subsequence $\{\mu_{n_{k(j)}}\}$ converging weakly $(j \to \infty)$ to some limit, and that limit must by hypothesis be μ. Thus every subsequence $\{\mu_{n_k}\}$ contains a further subsequence $\{\mu_{n_{k(j)}}\}$ converging weakly to μ.

Suppose that $\mu_n \Rightarrow \mu$ is false. Then there exists some x such that $\mu\{x\} = 0$ but $\mu_n(-\infty, x]$ does not converge to $\mu(-\infty, x]$. But then there exists a positive ϵ such that $|\mu_{n_k}(-\infty, x] - \mu(-\infty, x]| \geq \epsilon$ for an infinite sequence $\{n_k\}$ of integers, and no subsequence of $\{\mu_{n_k}\}$ can converge weakly to μ. This contradiction shows that $\mu_n \Rightarrow \mu$. ∎

If μ_n is a unit mass at x_n, then $\{\mu_n\}$ is tight if and only if $\{x_n\}$ is bounded. The theorem above and its corollary reduce in this case to standard facts about real line; see Example 25.4 and A10: tightness of sequences of probability measures is analogous to boundedness of sequences of real numbers.

Example 25.10. Let μ_n be the normal distribution with mean m_n and variance σ_n^2. If m_n and σ_n^2 are bounded, then the second moment of μ_n is bounded, and it follows by Markov's inequality (21.12) that $\{\mu_n\}$ is tight. The conclusion of Theorem 25.10 can also be checked directly: If $\{n_{k(j)}\}$ is chosen so that $\lim_j m_{n_{k(j)}} = m$ and $\lim_j \sigma^2_{n_{k(j)}} = \sigma^2$, then $\mu_{n_{k(j)}} \Rightarrow \mu$, where μ is normal with mean m and variance σ^2 (a unit mass at m if $\sigma^2 = 0$).

If $m_n > b$, then $\mu_n(b, \infty) \geq \frac{1}{2}$; if $m_n < a$, then $\mu_n(-\infty, a] \geq \frac{1}{2}$. Hence $\{\mu_n\}$ cannot be tight if m_n is unbounded. If m_n is bounded, say by K, then $\mu_n(-\infty, a] \geq \nu(-\infty, (a - K)\sigma_n^{-1}]$, where ν is the standard normal distribution. If σ_n is unbounded, then $\nu(-\infty, (a - K)\sigma_n^{-1}] \to \frac{1}{2}$ along some subsequence, and $\{\mu_n\}$ cannot be tight. Thus a sequence of normal distributions is tight if and only if the means and variances are bounded. ∎

Integration to the Limit

Theorem 25.11. *If* $X_n \Rightarrow X$, *then* $E[|X|] \leq \liminf_n E[|X_n|]$.

PROOF. Apply Skorohod's Theorem 25.6 to the distributions of X_n and X: There exist on a common probability space random variables Y_n and Y such that $Y = \lim_n Y_n$ with probability 1, Y_n has the distribution of X_n, and Y has the distribution of X. By Fatou's lemma, $E[|Y|] \leq \liminf_n E[|Y_n|]$. Since $|X|$ and $|Y|$ have the same distribution, they have the same expected value (see (21.6)), and similarly for $|X_n|$ and $|Y_n|$. ∎

The random variables X_n are said to be *uniformly integrable* if

$$(25.10) \qquad \lim_{\alpha \to \infty} \sup_n \int_{[|X_n| \geq \alpha]} |X_n| \, dP = 0;$$

see (16.21). This implies (see (16.22)) that

$$(25.11) \qquad \sup_n E[|X_n|] < \infty.$$

Theorem 25.12. *If* $X_n \Rightarrow X$ *and the* X_n *are uniformly integrable, then* X *is integrable and*

$$(25.12) \qquad E[X_n] \to E[X].$$

PROOF. Construct random variables Y_n and Y as in the preceding proof. Since $Y_n \to Y$ with probability 1 and the Y_n are uniformly integrable in the sense of (16.21), $E[X_n] = E[Y_n] \to E[Y] = E[X]$ by Theorem 16.14. ∎

If $\sup_n E[|X_n|^{1+\epsilon}] < \infty$ for some positive ϵ, then the X_n are uniformly integrable because

$$(25.13) \qquad \int_{[|X_n| \geq \alpha]} |X_n| \, dP \leq \frac{1}{\alpha^\epsilon} E[|X_n|^{1+\epsilon}].$$

Since $X_n \Rightarrow X$ implies that $X_n^r \Rightarrow X^r$ by Theorem 25.7, there is the following consequence of the theorem.

Corollary. *Let* r *be a positive integer. If* $X_n \Rightarrow X$ *and* $\sup_n E[|X_n|^{r+\epsilon}] < \infty$, *where* $\epsilon > 0$, *then* $E[|X|^r] < \infty$ *and* $E[X_n^r] \to E[X^r]$.

The X_n are also uniformly integrable if there is an integrable random variable Z such that $P[|X_n| \geq t] \leq P[|Z| \geq t]$ for $t > 0$, because then (21.10)

gives

$$\int_{[|X_n|\geq\alpha]} |X_n|\, dP \leq \int_{[|Z|\geq\alpha]} |Z|\, dP.$$

From this the dominated convergence theorem follows again.

PROBLEMS

25.1. (a) Show by example that distribution functions having densities can converge weakly even if the densities do not converge: *Hint:* Consider $f_n(x) = 1 + \cos 2\pi nx$ on $[0, 1]$.

(b) Let f_n be 2^n times the indicator of the set of x in the unit interval for which $d_{n+1}(x) = \cdots = d_{2n}(x) = 0$, where $d_k(x)$ is the kth dyadic digit. Show that $f_n(x) \to 0$ except on a set of Lebesgue measure 0; on this exceptional set, redefine $f_n(x) = 0$ for all n, so that $f_n(x) \to 0$ everywhere. Show that the distributions corresponding to these densities converge weakly to Lebesgue measure confined to the unit interval.

(c) Show that distributions with densities can converge weakly to a limit that has no density (even to a unit mass).

(d) Show that discrete distributions can converge weakly to a distribution that has a density.

(e) Construct an example, like that of Example 25.3, in which $\mu_n(A) \to \mu(A)$ fails but in which all the measures come from continuous densities on $[0, 1]$.

25.2. 14.8↑   Give a simple proof of the Gilvenko–Cantelli theorem (Theorem 20.6) under the extra hypothesis that F is continuous.

25.3. *Initial digits.* (a) Show that the first significant digit of a positive number x is d (in the scale of 10) if and only if $\{\log_{10} x\}$ lies between $\log_{10} d$ and $\log_{10}(d + 1)$, $d = 1, \ldots, 9$, where the braces denote fractional part.

(b) For positive numbers $x_1, x_2, \ldots$, let $N_r(d)$ be the number among the first n that have initial digit d. Show that

$$(25.14) \qquad \lim_n \frac{1}{n} N_n(d) = \log_{10}(d + 1) - \log_{10} d, \qquad d = 1, \ldots, 9,$$

if the sequence $\log_{10} x_n$, $n = 1, 2, \ldots$, is uniformly distributed modulo 1. This is true, for example, of $x_n = \vartheta^n$ if $\log_{10} \vartheta$ is irrational.

(c) Let D_n be the first significant digit of a positive random variable X_n. Show that

$$(25.15) \qquad \lim_n P[D_n = d] = \log_{10}(d + 1) - \log_{10} d, \qquad d = 1, \ldots, 9,$$

if $\{\log_{10} X_n\} \Rightarrow U$, where U is uniformly distributed over the unit interval.

25.4. Show that for each probability measure μ on the line there exist probability measures μ_n with finite support such that $\mu_n \Rightarrow \mu$. Show further that $\mu_n\{x\}$ can

be taken rational and that each point in the support can be taken rational. Thus there exists a countable set of probability measures such that every μ is the weak limit of some sequence from the set. The space of distribution functions is thus separable in the Lévy metric (see Problem 14.5).

25.5. Show that (25.5) implies that $P([X \leq x] \vartriangle [X_n \leq x]) \to 0$ if $P[X = x] = 0$.

25.6. For arbitrary random variables X_n there exist positive constants a_n such that $a_n X_n \Rightarrow 0$.

25.7. Generalize Example 25.8 by showing for three-dimensional random vectors (A_n, B_n, X_n) and constants a and b, $a \geq 0$, that, if $A_n \Rightarrow a$, $B_n \Rightarrow b$, and $X_n \Rightarrow X$, then $A_n X_n + B_n \Rightarrow aX + b$ *Hint:* First show that if $Y_n \Rightarrow Y$ and $D_n \Rightarrow 0$, then $D_n Y_n \Rightarrow 0$.

25.8. Suppose that $X_n \Rightarrow X$ and that h_n and h are Borel functions. Let E be the set of x for which $h_n x_n \to hx$ fails for some sequence $x_n \to x$. Suppose that $E \in \mathscr{R}^1$ and $P[X \in E] = 0$. Show that $h_n X_n \Rightarrow hX$.

25.9. Suppose that the distributions of random variables X_n and X have densities f_n and f. Show that if $f_n(x) \to f(x)$ for x outside a set of Lebesgue measure 0, then $X_n \Rightarrow X$.

25.10. ↑ Suppose that X_n assumes as values $\gamma_n + k\delta_n$, $k = 0, \pm 1, \ldots$, where $\delta_n > 0$. Suppose that $\delta_n \to 0$ and that, if k_n is an integer varying with n in such a way that $\gamma_n + k_n \delta_n \to x$, then $P[X_n = \gamma_n + k_n \delta_n] \delta_n^{-1} \to f(x)$, where f is the density of a random variable X. Show that $X_n \Rightarrow X$.

25.11. ↑ Let S_n have the binomial distribution with parameters n and p. Assume as known that

$$(25.16) \qquad P[S_n = k_n](np(1-p))^{1/2} \to \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

if $(k_n - np)(np(1-p))^{-1/2} \to x$. Deduce the DeMoivre–Laplace theorem: $(S_n - np)(np(1-p))^{-1/2} \Rightarrow N$, where N has the standard normal distribution. This is a special case of the central limit theorem; see Section 27.

25.12. Prove weak convergence in Example 25.3 by using Theorem 25.8 and the theory of the Riemann integral.

25.13. (a) Show that probability measures satisfy $\mu_n \Rightarrow \mu$ if $\mu_n(a, b] \to \mu(a, b]$ whenever $\mu\{a\} = \mu\{b\} = 0$.

(b) Show that, if $\int f \, d\mu_n \to \int f \, d\mu$ for all continuous f with bounded support, then $\mu_n \Rightarrow \mu$.

25.14. ↑ Let μ be Lebesgue measure confined to the unit interval; let μ_n correspond to a mass of $x_{n,i} - x_{n,i-1}$ at some point in $(x_{n,i-1}, x_{n,i}]$, where $0 = x_{n0}$ $< x_{n1} < \cdots < x_{nn} = 1$. Show by considering the distribution functions that $\mu_n \Rightarrow \mu$ if $\max_{i < n}(x_{n,i} - x_{n,i-1}) \to 0$. Deduce that a bounded Borel function continuous almost everywhere on the unit interval is Riemann integrable. See Problem 17.1.

25.15. 2.18 5.19↑ A function f of positive integers has *distribution function* F if F is the weak limit of the distribution function $P_n[m: f(m) \le x]$ of f under the measure having probability $1/n$ at each of $1, \ldots, n$ (see 2.34)). In this case $D[m: f(m) \le x] = F(x)$ (see (2.35)) for continuity points x of F. Show that $\varphi(m)/m$ (see (2.37)) has a distribution:

(a) Show by the mapping theorem that it suffices to prove that $f(m) = \log(\varphi(m)/m) = \sum_p \delta_p(m) \log(1 - 1/p)$ has a distribution.

(b) Let $f_u(m) = \sum_{p \le u} \delta_p(m) \log(1 - 1/p)$, and show by (5.45) that f_u has distribution function $F_u(x) = P[\sum_{p \le u} X_p \log(1 - 1/p) \le x]$, where the X_p are independent random variables (one for each prime p) such that $P[X_p = 1] = 1/p$ and $P[X_p = 0] = 1 - 1/p$.

(c) Show that $\sum_p X_p \log(1 - 1/p)$ converges with probability 1. *Hint:* Use Theorem 22.6.

(d) Show that $\lim_{u \to \infty} \sup_n E_n[|f - f_u|] = 0$ (see (5.46) for the notation).

(e) Conclude by Markov's inequality and Theorem 25.5 that f has the distribution of the sum in (c).

25.16. For $A \in \mathscr{R}^1$ and $T > 0$, put $\lambda_T(A) = \lambda([-T, T] \cap A)/2T$, where λ is Lebesgue measure. The *relative measure* of A is

$$(25.17) \qquad\qquad \rho(A) = \lim_{T \to \infty} \lambda_T(A),$$

provided that this limit exists. This is a continuous analogue of density (see (2.35)) for sets of integers. A Borel function f has a distribution under λ_T; if this converges weakly to F, then

$$(25.18) \qquad\qquad \rho[x: f(x) \le u] = F(u)$$

for continuity points u of F, and F is called the distribution function of f. Show that all periodic functions have distributions.

25.17. Suppose that $\sup_n \int f \, d\mu_n < \infty$ for a nonnegative f such that $f(x) \to \infty$ as $x \to \pm\infty$. Show that $\{\mu_n\}$ is tight.

25.18. 23.4↑ Show that the random variables A_t and L_t in Problems 23.3 and 23.4 converge in distribution. Show that the moments converge.

25.19. In the applications of Theorem 9.2, only a weaker result is actually needed: For each K there exists a positive $\alpha = \alpha(K)$ such that if $E[X] = 0$, $E[X^2] = 1$, and $E[X^4] \le K$, then $P[X \ge 0] \ge \alpha$. Prove this by using tightness and the corollary to Theorem 25.12.

25.20. Find uniformly integrable random variables X_n for which there is no integrable Z satisfying $P[|X_n| \ge t] \le P[|Z| \ge t]$ for $t > 0$.

SECTION 26. CHARACTERISTIC FUNCTIONS

Definition

The characteristic function of a probability measure μ on the line is defined for real t by

$$\varphi(t) = \int_{-\infty}^{\infty} e^{itx} \mu(dx)$$

$$= \int_{-\infty}^{\infty} \cos tx \, \mu(dx) + i \int_{-\infty}^{\infty} \sin tx \, \mu(dx);$$

see the end of Section 16 for integrals of complex-valued functions.[†] A random variable X with distribution μ has characteristic function

$$\varphi(t) = E[e^{itX}] = \int_{-\infty}^{\infty} e^{itx} \mu(dx).$$

The characteristic function is thus defined as the moment generating function but with the real argument s replaced by it; it has the advantage that it always exists because e^{itx} is bounded. The characteristic function in nonprobabilistic contexts is called the *Fourier transform*.

The characteristic function has three fundamental properties to be established here:

(i) If μ_1 and μ_2 have respective characteristic functions $\varphi_1(t)$ and $\varphi_2(t)$, then $\mu_1 * \mu_2$ has characteristic function $\varphi_1(t)\varphi_2(t)$. Although convolution is essential to the study of sums of independent random variables, it is a complicated operation, and it is often simpler to study the products of the corresponding characteristic functions.

(ii) The characteristic function uniquely determines the distribution. This shows that in studying the products in (i), no information is lost.

(iii) From the pointwise convergence of characteristic functions follows the weak convergence of the corresponding distributions. This makes it possible, for example, to investigate the asymptotic distributions of sums of independent random variables by means of their characteristic functions.

Moments and Derivatives

It is convenient first to study the relation between a characteristic function and the moments of the distribution it comes from.

[†] From complex variable theory only De Moivre's formula and the simplest properties of the exponential function are needed here.

Of course, $\varphi(0) = 1$, and by (16.30), $|\varphi(t)| \leq 1$ for all t. By Theorem 16.8(i), $\varphi(t)$ is continuous in t. In fact, $|\varphi(t+h) - \varphi(t)| \leq \int |e^{ihx} - 1| \mu(dx)$, and so it follows by the bounded convergence theorem that $\varphi(t)$ *is uniformly continuous.*

In the following relations, versions of Taylor's formula with remainder, x is assumed real. Integration by parts shows that

$$(26.1) \qquad \int_0^x (x-s)^n e^{is} \, ds = \frac{x^{n+1}}{n+1} + \frac{i}{n+1} \int_0^x (x-s)^{n+1} e^{is} \, ds,$$

and it follows by induction that

$$(26.2) \qquad e^{ix} = \sum_{k=0}^n \frac{(ix)^k}{k!} + \frac{i^{n+1}}{n!} \int_0^x (x-s)^n e^{is} \, ds$$

for $n \geq 0$. Replace n by $n-1$ in (26.1), solve for the integral on the right, and substitute this for the integral in (26.2); this gives

$$(26.3) \qquad e^{ix} = \sum_{k=0}^n \frac{(ix)^k}{k!} + \frac{i^n}{(n-1)!} \int_0^x (x-s)^{n-1} (e^{is} - 1) \, ds.$$

Estimating the integrals in (26.2) and (26.3) (consider separately the cases $x \geq 0$ and $x < 0$) now leads to

$$(26.4) \qquad \left| e^{ix} - \sum_{k=0}^n \frac{(ix)^k}{k!} \right| \leq \min\left\{ \frac{|x|^{n+1}}{(n+1)!}, \frac{2|x|^n}{n!} \right\}$$

for $n \geq 0$. The first term on the right gives a sharp estimate for $|x|$ small, the second a sharp estimate for $|x|$ large. For $n = 0, 1, 2$, the inequality specializes to

$$(26.4_0) \qquad\qquad\qquad |e^{ix} - 1| \leq \min\{|x|, 2\},$$

$$(26.4_1) \qquad\qquad\qquad |e^{ix} - (1+ix)| \leq \min\{\tfrac{1}{2}x^2, 2|x|\},$$

$$(26.4_2) \qquad\qquad\qquad \left| e^{ix} - (1 + ix - \tfrac{1}{2}x^2) \right| \leq \min\{\tfrac{1}{6}|x|^3, x^2\}.$$

If X has a moment of order n, it follows that

$$(26.5) \qquad \left| \varphi(t) - \sum_{k=0}^n \frac{(it)^k}{k!} E[X^k] \right| \leq E\left[\min\left(\frac{|tX|^{n+1}}{(n+1)!}, \frac{2|tX|^n}{n!} \right) \right].$$

For any t satisfying

(26.6)
$$\lim_n \frac{|t|^n E[|X|^n]}{n!} = 0,$$

$\varphi(t)$ must therefore have the expansion

(26.7)
$$\varphi(t) = \sum_{k=0}^{\infty} \frac{(it)^k}{k!} E[X^k];$$

compare (21.22). If

$$\sum_{k=0}^{\infty} \frac{|t|^k}{k!} E[|X|^k] = E[e^{|t\,X|}] < \infty,$$

then (see (16.31)) (26.7) must hold. Thus (26.7) holds if X has a moment generating function over the whole line.

Example 26.1. Since $E[e^{|t\,X|}] < \infty$ if X has the standard normal distribution, by (26.7) and (21.7) its characteristic function is

(26.8)
$$\varphi(t) = \sum_{k=0}^{\infty} \frac{(it)^{2k}}{(2k)!} 1 \times 3 \times \cdots \times (2k-1) = \sum_{k=0}^{\infty} \frac{1}{k!}\left(-\frac{t^2}{2}\right)^k = e^{-t^2/2}.$$

This and (21.25) formally coincide if $s = it$.   ∎

If the power-series expansion (26.7) holds, the moments of X can be read off from it:

(26.9)
$$\varphi^{(k)}(0) = i^k E[X^k].$$

This is the analogue of (21.23). It holds, however, under the weakest possible assumption, namely that $E[|X^k|] < \infty$. Indeed,

$$\frac{\varphi(t+h) - \varphi(t)}{h} - E[iXe^{it\,X}] = E\left[e^{it\,X}\frac{e^{ihX} - 1 - ihX}{h}\right].$$

By (26.4_1), the integrand on the right is dominated by $2|X|$ and goes to 0 with h; hence the expected value goes to 0 by the dominated convergence theorem. Thus $\varphi'(t) = E[iXe^{it\,X}]$. Repeating this argument inductively gives

(26.10)
$$\varphi^{(k)}(t) = E\left[(iX)^k e^{it\,X}\right]$$

if $E[|X^k|] < \infty$. Hence (26.9) holds if $E[|X^k|] < \infty$. The proof of uniform continuity for $\varphi(t)$ works for $\varphi^{(k)}(t)$ as well.

If $E[X^2]$ is finite, then

$$(26.11) \qquad \varphi(t) = 1 + itE[X] - \tfrac{1}{2}t^2E[X^2] + o(t^2), \qquad t \to 0.$$

Indeed, by (26.4$_2$), the error is at most $t^2E[\min\{|t||X|^3, X^2\}]$, and as $t \to 0$ the integrand goes to 0 and is dominated by X^2. Estimates of this kind are essential for proving limit theorems.

The more moments μ has, the more derivatives φ has. This is one sense in which lightness of the tails of μ is reflected by smoothness of φ. There are results which connect the behavior of $\varphi(t)$ as $|t| \to \infty$ with smoothness properties of μ. The *Riemann–Lebesgue theorem* is the most important of these:

Theorem 26.1. *If μ has a density, then $\varphi(t) \to 0$ as $|t| \to \infty$.*

PROOF. The problem is to prove for integrable f that $\int f(x)e^{itx}\,dx \to 0$ as $|t| \to \infty$. There exists by Theorem 17.1 a step function $g = \Sigma_k \alpha_k I_{A_k}$, a finite linear combination of indicators of intervals $A_k = (a_k, b_k]$, for which $\int|f - g|\,dx < \epsilon$. Now $\int f(x)e^{itx}\,dx$ differs by at most ϵ from $\int g(x)e^{itx}\,dx = \Sigma_k \alpha_k(e^{itb_k} - e^{ita_k})/it$, and this goes to 0 as $|t| \to \infty$. ∎

Independence

The multiplicative property (21.28) of moment generating functions extends to characteristic functions. Suppose that X_1 and X_2 are independent random variables with characteristic functions φ_1 and φ_2. If $Y_j = \cos X_j$ and $Z_j = \sin tX_j$, then (Y_1, Z_1) and (Y_2, Z_2) are independent; by the rules for integrating complex-valued functions,

$$
\begin{aligned}
\varphi_1(t)\varphi_2(t) &= (E[Y_1] + iE[Z_1])(E[Y_2] + iE[Z_2]) \\
&= E[Y_1]E[Y_2] - E[Z_1]E[Z_2] \\
&\quad + i(E[Y_1]E[Z_2] + E[Z_1]E[Y_2]) \\
&= E[Y_1Y_2 - Z_1Z_2 + i(Y_1Z_2 + Z_1Y_2)] = E[e^{it(X_1+X_2)}].
\end{aligned}
$$

This extends to sums of three or more: If $X_1, \ldots, X_n$ are independent, then

$$(26.12) \qquad E\left[e^{it\Sigma_{k=1}^n X_k}\right] = \prod_{k=1}^n E\left[e^{itX_k}\right].$$

If X has characteristic function $\varphi(t)$, then $aX + b$ has characteristic function

$$(26.13) \qquad E\left[e^{It(aX+b)}\right] = e^{itb}\varphi(at).$$

In particular, $-X$ has characteristic function $\varphi(-t)$, which is the complex conjugate of $\varphi(t)$.

Inversion and the Uniqueness Theorem

A characteristic function φ uniquely determines the measure μ it comes from. This fundamental fact will be derived by means of an inversion formula through which μ can in principle be recovered from φ.

Define

$$S(T) = \int_0^T \frac{\sin x}{x}\, dx, \qquad T \geq 0.$$

In Example 18.4 it is shown that

$$(26.14) \qquad \lim_{T \to \infty} S(T) = \frac{\pi}{2};$$

$S(T)$ is therefore bounded. If $\operatorname{sgn}\theta$ is $+1$, 0, or -1 as θ is positive, 0, or negative, then

$$(26.15) \qquad \int_0^T \frac{\sin t\theta}{t}\, dt = \operatorname{sgn}\theta \cdot S(T|\theta|), \qquad T \geq 0.$$

Theorem 26.2. *If the probability measure μ has characteristic function φ, and if $\mu\{a\} = \mu\{b\} = 0$, then*

$$(26.16) \qquad \mu(a,b] = \lim_{T \to \infty} \frac{1}{2\pi} \int_{-T}^T \frac{e^{-ita} - e^{-itb}}{it}\varphi(t)\, dt.$$

Distinct measures cannot have the same characteristic function.

Note: By (26.4_1) the integrand here converges as $t \to 0$ to $b - a$, which is to be taken as its value for $t = 0$. For fixed a and b the integrand is thus continuous in t, and by (26.4_0) it is bounded. If μ is a unit mass at 0, then $\varphi(t) \equiv 1$ and the integral in (26.16) cannot be extended over the whole line.

PROOF. The inversion formula will imply uniqueness: It will imply that if μ and ν have the same characteristic function, then $\mu(a,b] = \nu(a,b]$ if $\mu\{a\} = \nu\{a\} = \mu\{b\} = \nu\{b\} = 0$; but such intervals $(a,b]$ form a π-system generating $\mathscr{R}^1$.

Denote by I_T the quantity inside the limit in (26.16). By Fubini's theorem

$$(26.17) \qquad I_T = \frac{1}{2\pi} \int_{-\infty}^{\infty} \left[\int_{-T}^{T} \frac{e^{it(x-a)} - e^{it(x-b)}}{it} \, dt \right] \mu(dx).$$

This interchange is legitimate because the double integral extends over a set of finite product measure and by (26.4₀) the integrand is bounded by $|b - a|$. Rewrite the integrand by DeMoivre's formula. Since $\sin s$ and $\cos s$ are odd and even, respectively, (26.15) gives

$$I_T = \int_{-\infty}^{\infty} \left[\frac{\text{sgn}(x - a)}{\pi} S(T \cdot |x - a|) - \frac{\text{sgn}(x - b)}{\pi} S(T \cdot |x - b|) \right] \mu(dx).$$

The integrand here is bounded and converges as $T \to \infty$ to the function

$$(26.18) \qquad \psi_{a,b}(x) = \begin{cases} 0 & \text{if } x < a, \\ \frac{1}{2} & \text{if } x = a, \\ 1 & \text{if } a < x < b, \\ \frac{1}{2} & \text{if } x = b, \\ 0 & \text{if } b < x. \end{cases}$$

Thus $I_T \to \int \psi_{a,b} \, d\mu$, which implies that (26.16) holds if $\mu\{a\} = \mu\{b\} = 0$. ∎
The inversion formula contains further information. Suppose that

$$(26.19) \qquad \int_{-\infty}^{\infty} |\varphi(t)| \, dt < \infty.$$

In this case the integral in (26.16) can be extended over R^1. By (26.4₀),

$$\left| \frac{e^{-itb} - e^{-ita}}{it} \right| = \frac{|e^{it(b-a)} - 1|}{|t|} \leq |b - a|;$$

therefore, $\mu(a, b) \leq (b - a)\int_{-\infty}^{\infty} |\varphi(t)| \, dt$, and there can be no point masses. By (26.16), the corresponding distribution function satisfies

$$\frac{F(x + h) - F(x)}{h} = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{e^{-itx} - e^{-it(x+h)}}{ith} \varphi(t) \, dt$$

(whether h is positive or negative). The integrand is by (26.4₀) dominated by $|\varphi(t)|$ and goes to $e^{itx}\varphi(t)$ as $h \to 0$. Therefore, F has derivative

$$(26.20) \qquad f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \varphi(t) \, dt.$$

Since f is continuous for the same reason φ is, it integrates to F by the fundamental theorem of the calculus (see (17.6)). Thus (26.19) *implies that μ has the continuous density* (26.20). Moreover, this is the only continuous density. In this result, as in the Riemann–Lebesgue theorem, conditions on the size of $\varphi(t)$ for large $|t|$ are connected with smoothness properties of μ.

The inversion formula (26.20) has many applications. In the first place, it can be used for a new derivation of (26.14). As pointed out in Example 17.3, the existence of the limit in (26.14) is easy to prove. Denote this limit temporarily by $\pi_0/2$—without assuming that $\pi_0 = \pi$. Then (26.16) and (26.20) follow as before if π is replaced by π_0. Applying the latter to the standard normal density (see (26.8)) gives

$$(26.21) \qquad \frac{1}{\sqrt{2\pi}} e^{-x^2/2} = \frac{1}{2\pi_0} \int_{-\infty}^{\infty} e^{-itx} e^{-t^2/2}\, dt,$$

where the π on the left is that of analysis and geometry—it comes ultimately from the quadrature (18.10). An application of (26.8) with x and t interchanged reduces the right side of (26.21) to $(\sqrt{2\pi}/2\pi_0)e^{-x^2/2}$, and therefore π_0 does equal π.

Consider the densities in the table. The characteristic function for the *normal* distribution has already been calculated. For the *uniform* distribution over $(0,1)$, the computation is of course straightforward; note that in this case the density cannot be recovered from (26.20), because $\varphi(t)$ is not integrable; this is reflected in the fact that the density has discontinuities at 0 and 1.

Distribution	Density	Interval	Characteristic Function
1. Normal	$\dfrac{1}{\sqrt{2\pi}} e^{-x^2/2}$	$-\infty < x < \infty$	$e^{-t^2/2}$
2. Uniform	1	$0 < x < 1$	$\dfrac{e^{it} - 1}{it}$
3. Exponential	e^{-x}	$0 < x < \infty$	$\dfrac{1}{1 - it}$
4. Double exponential or Laplace	$\tfrac{1}{2} e^{-|x|}$	$-\infty < x < \infty$	$\dfrac{1}{1 + t^2}$
5. Cauchy	$\dfrac{1}{\pi}\dfrac{1}{1 + x^2}$	$-\infty < x < \infty$	$e^{-|t|}$
6. Triangular	$1 - |x|$	$-1 < x < 1$	$2\dfrac{1 - \cos t}{t^2}$
7.	$\dfrac{1}{\pi}\dfrac{1 - \cos x}{x^2}$	$-\infty < x < \infty$	$(1 - |t|)I_{(-1,1)}(t)$

The characteristic function for the *exponential* distribution is easily calculated; compare Example 21.3. As for the *double exponential* or *Laplace* distribution, $e^{-|x|}e^{itx}$ integrates over $(0, \infty)$ to $(1 - it)^{-1}$ and over $(-\infty, 0)$ to $(1 + it)^{-1}$, which gives the result. By (26.20), then,

$$e^{-|x|} = \frac{1}{\pi} \int_{-\infty}^{\infty} e^{-itx} \frac{dt}{1 + t^2} .$$

For $x = 0$ this gives the standard integral $\int_{-\infty}^{\infty} dt / (1 + t^2) = \pi$; see Example 17.5. Thus the *Cauchy* density in the table integrates to 1 and has characteristic function $e^{-|t|}$. This distribution has no first moment, and the characteristic function is not differentiable at the origin.

A straightforward integration shows that the *triangular* density has the characteristic function given in the table, and by (26.20),

$$(1 - |x|) I_{(-1, 1)}(x) = \frac{1}{\pi} \int_{-\infty}^{\infty} e^{-itx} \frac{1 - \cos t}{t^2} \, dt.$$

For $x = 0$ this is $\int_{-\infty}^{\infty} (1 - \cos t) t^{-2} \, dt = \pi$; hence the last line of the table.

Each density and characteristic function in the table can be transformed by (26.13), which gives a family of distributions.

The Continuity Theorem

Because of (26.12), the characteristic function provides a powerful means of studying the distributions of sums of independent random variables. It is often easier to work with products of characteristic functions than with convolutions, and knowing the characteristic function of the sum is by Theorem 26.2 in principle the same thing as knowing the distribution itself. Because of the following *continuity theorem*, characteristic functions can be used to study limit distributions.

Theorem 26.3. *Let* μ_n, μ *be probability measures with characteristic functions* φ_n, φ. *A necessary and sufficient condition for* $\mu_n \Rightarrow \mu$ *is that* $\varphi_n(t) \to \varphi(t)$ *for each* t.

PROOF. *Necessity.* For each t, e^{itx} has bounded modulus and is continuous in x. The necessity therefore follows by an application of Theorem 25.8 (to the real and imaginary parts of e^{itx}).

Sufficiency. By Fubini's theorem,

$$(26.22) \qquad \frac{1}{u}\int_{-u}^{u}(1 - \varphi_n(t))\,dt = \int_{-\infty}^{\infty}\left[\frac{1}{u}\int_{-u}^{u}(1 - e^{itx})\,dt\right]\mu_n(dx)$$

$$= 2\int_{-\infty}^{\infty}\left(1 - \frac{\sin ux}{ux}\right)\mu_n(dx)$$

$$\geq 2\int_{|x|\geq 2/u}\left(1 - \frac{1}{|ux|}\right)\mu_n(dx)$$

$$\geq \mu_n\left[x : |x| \geq \frac{2}{u}\right].$$

(Note that the first integral is real.) Since φ is continuous at the origin and $\varphi(0) = 1$, there is for positive ϵ a u for which $u^{-1}\int_{-u}^{u}(1 - \varphi(t))\,dt < \epsilon$. Since φ_n converges to φ, the bounded convergence theorem implies that there exists an n_0 such that $u^{-1}\int_{-u}^{u}(1 - \varphi_n(t))\,dt < 2\epsilon$ for $n \geq n_0$. If $a = 2/u$ in (26.22), then $\mu_n[x : |x| \geq a] < 2\epsilon$ for $n \geq n_0$. Increasing a if necessary will ensure that this inequality also holds for the finitely many n preceding n_0. Therefore, $\{\mu_n\}$ is tight.

By the corollary to Theorem 25.10, $\mu_n \Rightarrow \mu$ will follow if it is shown that each subsequence $\{\mu_{n_k}\}$ that converges weakly at all converges weakly to μ. But if $\mu_{n_k} \Rightarrow \nu$ as $k \to \infty$, then by the necessity half of the theorem, already proved, ν has characteristic function $\lim_k \varphi_{n_k}(t) = \varphi(t)$. By Theorem 26.2, ν and μ must coincide.   ■

Two corollaries, interesting in themselves, will make clearer the structure of the proof of sufficiency given above. In each, let μ_n be probability measures on the line with characteristic functions φ_n.

Corollary 1. *Suppose that* $\lim_n \varphi_n(t) = g(t)$ *for each* t, *where the limit function g is continuous at* 0. *Then there exists a μ such that $\mu_n \Rightarrow \mu$, and μ has characteristic function g.*

PROOF. The point of the corollary is that g is not assumed at the outset to be a characteristic function. But in the argument following (26.22), only $\varphi(0) = 1$ and the continuity of φ at 0 were used; hence $\{\mu_n\}$ is tight under the present hypothesis. If $\mu_{n_k} \Rightarrow \nu$ as $k \to \infty$, then ν must have characteristic function $\lim_k \varphi_{n_k}(t) = g(t)$. Thus g is, in fact, a characteristic function, and the proof goes through as before.   ■

In this proof the continuity of g was used to establish tightness. Hence if $\{\mu_n\}$ is assumed tight in the first place, the hypothesis of continuity can be suppressed:

Corollary 2.   *Suppose that* $\lim_n \varphi_n(t) = g(t)$ *exists for each* t *and that* $\{\mu_n\}$ *is tight. Then there exists a* μ *such that* $\mu_n \Rightarrow \mu$, *and* μ *has characteristic function* g.

This second corollary applies, for example, if the μ_n have a common bounded support.

Example 26.2.   If μ_n is the uniform distribution over $(-n, n)$, its characteristic function is $(nt)^{-1} \sin tn$ for $t \neq 0$, and hence it converges to $I_{\{0\}}(t)$. In this case $\{\mu_n\}$ is not tight, the limit function is not continuous at 0, and μ_n does not converge weakly.   ∎

Fourier Series*

Let μ be a probability measure on $\mathscr{R}^1$ that is supported by $[0, 2\pi]$. Its Fourier coefficients are defined by

$$(26.23) \qquad c_m = \int_0^{2\pi} e^{imx} \mu(dx), \qquad m = 0, \pm 1, \pm 2, \ldots .$$

These coefficients, the values of the characteristic function for integer arguments, suffice to determine μ except for the weights it may put at 0 and 2π. The relation between μ and its Fourier coefficients can be expressed formally by

$$(26.24) \qquad \mu(dx) \sim \frac{1}{2\pi} \sum_{l = -\infty}^{\infty} c_l e^{-ilx} dx:$$

if the $\mu(dx)$ in (26.23) is replaced by the right side of (26.24), and if the sum over l is interchanged with the integral, the result is a formal identity.

To see how to recover μ from its Fourier coefficients, consider the symmetric partial sums $s_m(t) = (2\pi)^{-1} \sum_{l=-m}^{m} c_l e^{-ilt}$ and their Cesàro averages $\sigma_m(t) = m^{-1} \sum_{l=0}^{m-1} s_l(t)$. From the trigonometric identity [A24]

$$(26.25) \qquad \sum_{l=0}^{m-1} \sum_{k=-l}^{l} e^{ikx} = \frac{\sin^2 \frac{1}{2} mx}{\sin^2 \frac{1}{2} x}$$

it follows that

$$(26.26) \qquad \sigma_m(t) = \frac{1}{2\pi m} \int_0^{2\pi} \frac{\sin^2 \frac{1}{2} m(x-t)}{\sin^2 \frac{1}{2}(x-t)} \mu(dx).$$

*This topic may be omitted

If μ is $(2\pi)^{-1}$ times Lebesgue measure confined to $[0, 2\pi]$, then $c_0 = 1$ and $c_m = 0$ for $m \neq 0$, so that $\sigma_m(t) = s_m(t) = (2\pi)^{-1}$; this gives the identity

$$(26.27) \qquad \frac{1}{2\pi m} \int_{-\pi}^{\pi} \frac{\sin^2 \frac{1}{2} ms}{\sin^2 \frac{1}{2} s} \, ds = 1.$$

Suppose that $0 < a < b < 2\pi$, and integrate (26.26) over (a, b). Fubini's theorem (the integrand is nonnegative) and a change of variable lead to

$$(26.28) \qquad \int_a^b \sigma_m(t) \, dt = \int_0^{2\pi} \left[\frac{1}{2\pi m} \int_{a-x}^{b-x} \frac{x \sin^2 \frac{1}{2} ms}{\sin^2 \frac{1}{2} s} \, ds \right] \mu(dx).$$

The denominator in (26.27) is bounded away from 0 outside $(-\delta, \delta)$, and so as m goes to ∞ with δ fixed $(0 < \delta < \pi)$,

$$\frac{1}{2\pi m} \int_{\delta < |s| < \pi} \frac{\sin^2 \frac{1}{2} ms}{\sin^2 \frac{1}{2} s} \, ds \to 0, \qquad \frac{1}{2\pi m} \int_{|s| < \delta} \frac{\sin^2 \frac{1}{2} ms}{\sin^2 \frac{1}{2} s} \, ds \to 1.$$

Therefore, the expression in brackets in (26.28) goes to 0 if $0 \leq x < a$ or $b < x \leq 2\pi$, and it goes to 1 if $a < x < b$; and because of (26.27), it is bounded by 1. It follows by the bounded convergence theorem that

$$(26.29) \qquad \mu(a, b] = \lim_m \int_a^b \sigma_m(t) \, dt$$

if $\mu\{a\} = \mu\{b\} = 0$ and $0 < a < b < 2\pi$.

This is the analogue of (26.16). If μ and ν have the same Fourier coefficients, it follows from (26.29) that $\mu(A) = \nu(A)$ for $A \subset (0, 2\pi)$ and hence that $\mu\{0, 2\pi\} = \nu\{0, 2\pi\}$. It is clear from periodicity that the coefficients (26.23) are unchanged if $\mu\{0\}$ and $\mu\{2\pi\}$ are altered but $\mu\{0\} + \mu\{2\pi\}$ is held constant.

Suppose that μ_n is supported by $[0, 2\pi]$ and has coefficients $c_m^{(n)}$, and suppose that $\lim_n c_m^{(n)} = c_m$ for all m. Since $\{\mu_n\}$ is tight, $\mu_n \Rightarrow \mu$ will hold if $\mu_{n_k} \Rightarrow \nu$ $(k \to \infty)$ implies $\nu = \mu$. But in this case ν and μ have the same coefficients c_m, and hence they are identical except perhaps in the way they split the mass $\nu\{0, 2\pi\} = \mu\{0, 2\pi\}$ between the points 0 and 2π. But this poses no problem if $\mu\{0, 2\pi\} = 0$: *If* $\lim_n c_m^{(n)} = c_m$ *for all* m *and* $\mu\{0\} = \mu\{2\pi\} = 0$, *then* $\mu_n \Rightarrow \mu$.

Example 26.3. If μ is $(2\pi)^{-1}$ times Lebesgue measure confined to the interval $[0, 2\pi]$, the condition is that $\lim_n c_m^{(n)} = 0$ for $m \neq 0$. Let $x_1, x_2, \ldots$ be a sequence of reals, and let μ_n put mass n^{-1} at each point $2\pi\{x_k\}$, $1 \leq k \leq n$, where $\{x_k\} = x_k - \lfloor x_k \rfloor$ denotes fractional part. This is the probability measure (25.3) rescaled to $[0, 2\pi]$. The sequence $x_1, x_2, \ldots$ is uniformly distributed modulo 1 if and only if

$$\frac{1}{n} \sum_{k=1}^n e^{2\pi i \{x_k\} m} = \frac{1}{n} \sum_{k=1}^n e^{2\pi i x_k m} \to 0$$

for $m \neq 0$. This is *Weyl's criterion*.

If $x_k = k\theta$, where θ is irrational, then $\exp(2\pi i\theta m) \neq 1$ for $m \neq 0$ and hence

$$\frac{1}{n} \sum_{k=1}^{n} e^{2\pi i k\theta m} = \frac{1}{n} e^{2\pi i\theta m} \frac{1 - e^{2\pi i n\theta m}}{1 - e^{2\pi i\theta m}} \to 0.$$

Thus $\theta, 2\theta, 3\theta, \ldots$ is uniformly distributed modulo 1 if θ is irrational, which gives another proof of Theorem 25.1.   ■

PROBLEMS

26.1. A random variable has a *lattice distribution* if for some a and b, $b > 0$, the lattice $[a + nb:\ n = 0, \pm 1, \ldots]$ supports the distribution of X. Let X have characteristic function φ.

(a) Show that a necessary condition for X to have a lattice distribution is that $|\varphi(t)| = 1$ for some $t \neq 0$.

(b) Show that the condition is sufficient as well.

(c) Suppose that $|\varphi(t)| = |\varphi(t')| = 1$ for incommensurable t and t' ($t \neq 0$, $t' \neq 0$, t/t' irrational). Show that $P[X = c] = 1$ for some constant c.

26.2. If $\mu(-\infty, x] = \mu[-x, \infty)$ for all x (which implies that $\mu(A) = \mu(-A)$ for all $A \in \mathscr{R}^1$), then μ is *symmetric*. Show that this holds if and only if the characteristic function is real.

26.3. Consider functions φ that are real and nonnegative and satisfy $\varphi(-t) = \varphi(t)$ and $\varphi(0) = 1$.

(a) Suppose that $d_1, d_2, \ldots$ are positive and $\sum_{k=1}^{\infty} d_k = \infty$, that $s_1 \geq s_2 \geq \cdots \geq 0$ and $\lim_k s_k = 0$, and that $\sum_{k=1}^{\infty} s_k d_k = 1$. Let φ be the convex polygon whose successive sides have slopes $-s_1, -s_2, \ldots$ and lengths $d_1, d_2, \ldots$ when projected on the horizontal axis: φ has value $1 - \sum_{j=1}^{k} s_j d_j$ at $t_k = d_1 + \cdots + d_k$. If $s_n = 0$, there are in effect only n sides. Let $\varphi_0(t) = (1 - |t|)I_{(-1,1)}(t)$ be the characteristic function in the last line in the table on p. 348, and show that $\varphi(t)$ is a convex combination of the characteristic functions $\varphi_0(t/t_k)$ and hence is itself a characteristic function.

(b) *Pólya's criterion.* Show that φ is a characteristic function if it is even and continuous and, on $[0, \infty)$, nonincreasing and convex ($\varphi(0) = 1$).
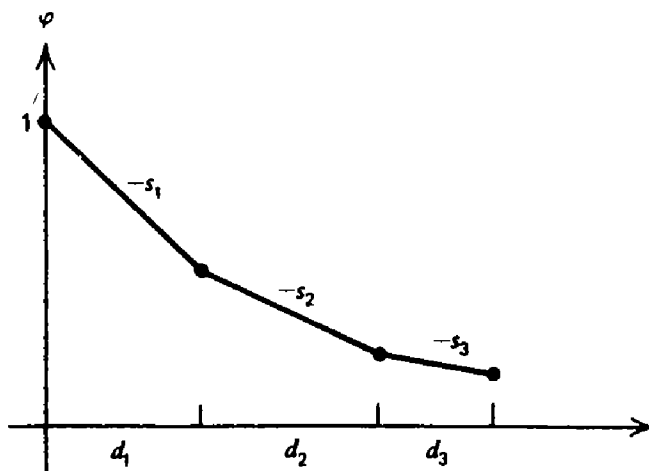
26.4. ↑ Let φ_1 and φ_2 be characteristic functions, and show that the set $A = [t: \varphi_1(t) = \varphi_2(t)]$ is closed, contains 0, and is symmetric about 0. Show that every set with these three properties can be such an A. What does this say about the uniqueness theorem?

26.5. Show by Theorem 26.1 and integration by parts that if μ has a density f with integrable derivative f', then $\varphi(t) = o(t^{-1})$ as $|t| \to \infty$. Extend to higher derivatives.

26.6. Show for independent random variables uniformly distributed over $(-1, +1)$ that $X_1 + \cdots + X_n$ has density $\pi^{-1}\int_0^\infty ((\sin t)/t)^n \cos tx\, dt$ for $n \geq 2$.

26.7. 21.17↑ *Uniqueness theorem for moment generating functions.* Suppose that F has a moment generating function in $(-s_0, s_0)$, $s_0 > 0$. From the fact that $\int_{-\infty}^\infty e^{zx}\, dF(x)$ is analytic in the strip $-s_0 < \operatorname{Re} z < s_0$, prove that the moment generating function determines F. Show that it is enough that the moment generating function exist in $[0, s_0)$, $s_0 > 0$.

26.8. 21.20   26.7↑ Show that the gamma density (20.47) has characteristic function

$$\frac{1}{(1 - it/\alpha)^u} = \exp\left[-u \log\left(1 - \frac{it}{\alpha}\right)\right],$$

where the logarithm is the principal part. Show that $\int_0^\infty e^{zx} f(x; \alpha, u)\, dx$ is analytic for $\operatorname{Re} z < \alpha$.

26.9. Use characteristic functions for a simple proof that the family of Cauch distributions defined by (20.45) is closed under convolution; compare th argument in Problem 20.14(a). Do the same for the normal distributic (compare Example 20.6) and for the Poisson and gamma distributions.

26.10. Suppose that $F_n \Rightarrow F$ and that the characteristic functions are dominated by integrable function. Show that F has a density that is the limit of the densit of the F_n.

26.11. Show for all a and b that the right side of (26.16) is $\mu(a, b) + \frac{1}{2}\mu\{a\} + \frac{1}{2}\mu$

26.12. By the kind of argument leading to (26.16), show that

$$(26.30) \qquad \mu\{a\} = \lim_{T\to\infty} \frac{1}{2T}\int_{-T}^T e^{-ita}\varphi(t)\, dt.$$

26.13. ↑ Let $x_1, x_2, \ldots$ be the points of positive μ-measure. By the following prove that

$$(26.31) \qquad \lim_{T\to\infty} \frac{1}{2T}\int_{-T}^T |\varphi(t)|^2\, dt = \sum_k (\mu\{x_k\})^2.$$

Let X and Y be independent and have characteristic function φ.

(a) Show by (26.30) that the left side of (26.31) is $P[X - Y = 0]$.

(b) Show (Theorem 20.3) that $P[X - Y = 0] = \int_{-\infty}^{\infty} P[X = y]\mu(dy) = \sum_k (\mu\{x_k\})^2$.

26.14. ↑ Show that μ has no point masses if $\varphi^2(t)$ is integrable.

26.15. (a) Show that if $\{\mu_n\}$ is tight, then the characteristic functions $\varphi_n(t)$ are uniformly equicontinuous (for each ϵ there is a δ such that $|s - t| < \delta$ implies that $|\varphi_n(s) - \varphi_n(t)| < \epsilon$ for all n).

(b) Show that $\mu_n \Rightarrow \mu$ implies that $\varphi_n(t) \to \varphi(t)$ uniformly on bounded sets.

(c) Show that the convergence in part (b) need not be uniform over the entire line.

26.16. 14.5 26.15 ↑ For distribution functions F and G, define $d'(F, G) = \sup_t |\varphi(t) - \psi(t)|/(1 + |t|)$, where φ and ψ are the corresponding characteristic functions. Show that this is a metric and equivalent to the Lévy metric.

26.17. 25.16 ↑ A real function f has *mean value*

$$(26.32) \qquad M[f(x)] = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} f(x)\, dx,$$

provided that f is integrable over each $[-T, T]$ and the limit exists.

(a) Show that, if f is bounded and $e^{itf(x)}$ has a mean value for each t, then f has a distribution in the sense of (25.18).

(b) Show that

$$(26.33) \qquad M[e^{itx}] = \begin{cases} 1 & \text{if } t = 0, \\ 0 & \text{if } t \neq 0. \end{cases}$$

Of course, $f(x) = x$ has no distribution.

26.18. Suppose that X is irrational with probability 1. Let μ_n be the distribution of the fractional part $\{nX\}$. Use the continuity theorem and Theorem 25.1 to show that $n^{-1}\sum_{k=1}^{n}\mu_k$ converges weakly to the uniform distribution on $[0, 1]$.

26.19. 25.13 ↑ The uniqueness theorem for characteristic functions can be derived from the Weierstrass approximation theorem. Fill in the details of the following argument. Let μ and ν be probability measures on the line. For continuous f with bounded support choose a so that $\mu(-a, a)$ and $\nu(-a, a)$ are nearly 1 and f vanishes outside $(-a, a)$. Let g be periodic and agree with f in $(-a, a)$, and by the Weierstrass theorem uniformly approximate $g(x)$ by a trigonometric sum $p(x) = \sum_{k=1}^{N} a_k e^{it_k x}$. If μ and ν have the same characteristic function, then $\int f\, d\mu \approx \int g\, d\mu \approx \int p\, d\mu = \int p\, d\nu \approx \int g\, d\nu \approx \int f\, d\nu$.

26.20. Use the continuity theorem to prove the result in Example 25.2 concerning the convergence of the binomial distribution to the Poisson.

26.21. According to Example 25.8, if $X_n \Rightarrow X$, $a_n \to a$, and $b_n \to b$, then $a_n X_n + b_n \Rightarrow aX + b$. Prove this by means of characteristic functions.

26.22. 26.1 26.15↑ According to Theorem 14.2, if $X_n \Rightarrow X$ and $a_n X_n + b_n \Rightarrow Y$, where $a_n > 0$ and the distributions of X and Y are nondegenerate, then $a_n \to a > 0$, $b_n \to b$, and $aX + b$ and Y have the same distribution. Prove this by characteristic functions. Let φ_n, φ, ψ be the characteristic functions of X_n, X, Y.

(a) Show that $|\varphi_n(a_n t)| \to |\psi(t)|$ uniformly on bounded sets and hence that a_n cannot converge to 0 along a subsequence.

(b) Interchange the roles of φ and ψ and show that a_n cannot converge to infinity along a subsequence.

(c) Show that a_n converges to some $a > 0$.

(d) Show that $e^{itb_n} \to \psi(t)/\varphi(at)$ in a neighborhood of 0 and hence that $\int_0^t e^{isb_n} ds \to \int_0^t (\psi(s)/\varphi(as)) ds$. Conclude that b_n converges.

26.23. Prove a continuity theorem for moment generating functions as defined by (22.4) for probability measures on $[0, \infty)$. For uniqueness, see Theorem 22.2; the analogue of (26.22) is

$$\frac{2}{u} \int_0^u (1 - M(s)) \, ds \ge \mu\left(\frac{2}{u}, \infty\right).$$

26.24. 26.4↑ Show by example that the values $\varphi(m)$ of the characteristic function at integer arguments may not determine the distribution if it is not supported by $[0, 2\pi]$.

26.25. If f is integrable over $[0, 2\pi]$, define its Fourier coefficients as $\int_0^{2\pi} e^{imx} f(x) \, dx$. Show that these coefficients uniquely determine f up to sets of measure 0.

26.26. 19.8 26.25↑ Show that the trigonometric system (19.17) is complete.

26.27. The Fourier-series analogue of the condition (26.19) is $\sum_m |c_m| < \infty$. Show that it implies μ has density $f(x) = (2\pi)^{-1} \sum_m c_m e^{-imx}$ on $[0, 2\pi]$, where f is continuous and $f(0) = f(2\pi)$. This is the analogue of the inversion formula (26.20).

26.28. ↑ Show that

$$(\pi - x)^2 = \frac{\pi^2}{3} + 4 \sum_{m=1}^{\infty} \frac{\cos mx}{m^2}, \qquad 0 \le x \le 2\pi.$$

Show that $\sum_{m=1}^{\infty} 1/m^2 = \pi^2/6$ and $\sum_{m=1}^{\infty} (-1)^{m+1}/m^2 = \pi^2/12$.

26.29. (a) Suppose X' and X'' are independent random variables with values in $[0, 2\pi]$, and let X be $X' + X''$ reduced module 2π. Show that the corresponding Fourier coefficients satisfy $c_m = c'_m c''_m$.

(b) Show that if one or the other of X' and X'' is uniformly distributed, so is X.

26.30. 26.25 ↑ The theory of Fourier series can be carried over from $[0, 2\pi]$ to the unit circle in the complex plane with normalized circular Lebesgue measure P. The circular functions e^{imx} become the powers ω^m, and an integrable f is determined to within sets of measure 0 by its Fourier coefficients $c_m = \int_\Omega \omega^m f(\omega) P(d\omega)$. Suppose that A is invariant under the rotation through the angle $\arg c$ (Example 24.4). Find a relation on the Fourier coefficients of I_A, and conclude that the rotation is ergodic if c is not a root of unity. Compare the proof on p. 316.

SECTION 27. THE CENTRAL LIMIT THEOREM

Identically Distributed Summands

The central limit theorem says roughly that the sum of many independent random variables will be approximately normally distributed if each summand has high probability of being small. Theorem 27.1, the *Lindeberg–Lévy theorem*, will give an idea of the techniques and hypotheses needed for the more general results that follow.

Throughout, N will denote a random variable with the standard normal distribution:

$$(27.1) \qquad P[N \in A] = \frac{1}{\sqrt{2\pi}} \int_A e^{-x^2/2} dx.$$

Theorem 27.1. *Suppose that* $\{X_n\}$ *is an independent sequence of random variables having the same distribution with mean* c *and finite positive variance* σ^2. *If* $S_n = X_1 + \cdots + X_n$, *then*

$$(27.2) \qquad \frac{S_n - nc}{\sigma\sqrt{n}} \Rightarrow N.$$

By the argument in Example 25.7, (27.2) implies that $n^{-1}S_n \Rightarrow c$. The central limit theorem and the strong law of large numbers thus refine the weak law of large numbers in different directions.

Since Theorem 27.1 is a special case of Theorem 27.2, no proof is really necessary. To understand the methods of this section, however, consider the special case in which X_k takes the values ± 1 with probability $1/2$ each. Each X_k then has characteristic function $\varphi(t) = \frac{1}{2}e^{it} + \frac{1}{2}e^{-it} = \cos t$. By (26.12) and (26.13), $S_n/\sqrt{n}$ has characteristic function $\varphi^n(t/\sqrt{n})$, and so, by the continuity theorem, the problem is to show that $\cos^n t/\sqrt{n} \to E[e^{itN}] = e^{-t^2/2}$, or that $n \log \cos t/\sqrt{n}$ (well defined for large n) goes to $-\frac{1}{2}t^2$. But this follows by l'Hopital's rule: Let $t/\sqrt{n} = x$ go continuously to 0.

For a proof closer in spirit to those that follow, note that (26.5) for $n = 2$ gives $|\varphi(t) - (1 - \frac{1}{2}t^2)| \leq |t|^3$ $(|X_k| \leq 1)$. Therefore,

$$(27.3) \qquad \left| \varphi\left(\frac{t}{\sqrt{n}}\right) - \left(1 - \frac{t^2}{2n}\right) \right| \leq \frac{|t|^3}{n^{3/2}}.$$

Rather than take logarithms, use (27.5) below, which gives (n large)

$$(27.4) \qquad \left| \varphi^n\left(\frac{t}{\sqrt{n}}\right) - \left(1 - \frac{t^2}{2n}\right)^n \right| \leq \frac{|t|^3}{\sqrt{n}} \to 0.$$

But of course $(1 - t^2/2n)^n \to e^{-t^2/2}$, which completes the proof for this special case.

Logarithms for complex arguments can be avoided by use of the following simple lemma.

Lemma 1. *Let* $z_1, \ldots, z_m$ *and* $w_1, \ldots, w_m$ *be complex numbers of modulus at most* 1; *then*

$$(27.5) \qquad |z_1 \cdots z_m - w_1 \cdots w_m| \leq \sum_{k=1}^{m} |z_k - w_k|.$$

PROOF. This follows by induction from $z_1 \cdots z_m - w_1 \cdots w_m = (z_1 - w_1)(z_2 \cdots z_m) + w_1(z_2 \cdots z_m - w_2 \cdots w_m)$. ∎

Two illustrations of Theorem 27.1:

Example 27.1. In the classical De Moivre-Laplace theorem, X_n takes the values 1 and 0 with probabilities p and $q = 1 - p$, so that $c = p$, and $\sigma^2 = pq$. Here S_n is the number of successes in n Bernoulli trials, and $(S_n - np)/\sqrt{npq} \Rightarrow N$. ∎

Example 27.2. Suppose that one wants to estimate the parameter α of an exponential distribution (20.10) on the basis of an independent sample $X_1, \ldots, X_n$. As $n \to \infty$ the sample mean $\bar{X}_n = n^{-1}\sum_{k=1}^{n}X_k$ converges in probability to the mean $1/\alpha$ of the distribution, and hence it is natural to use $1/\bar{X}_n$ to estimate α itself. How good is the estimate? The variance of the exponential distribution being $1/\alpha^2$ (Example 21.3), $\alpha\sqrt{n}\,(\bar{X}_n - 1/\alpha) \Rightarrow N$ by the Lindeberg-Lévy theorem. Thus $\bar{X}_n$ is approximately normally distributed with mean $1/\alpha$ and standard deviation $1/\alpha\sqrt{n}$.

By Skorohod's Theorem 25.6 there exist on a single probability space random variables $\bar{Y}_n$ and Y having the respective distributions of $\bar{X}_n$ and N and satisfying $\alpha\sqrt{n}\,(\bar{Y}_n(\omega) - 1/\alpha) \to Y(\omega)$ for each ω. Now $\bar{Y}_n(\omega) \to 1/\alpha$ and $\alpha^{-1}\sqrt{n}\,(\bar{Y}_n(\omega)^{-1} - \alpha) = \alpha\sqrt{n}\,(\alpha^{-1} - \bar{Y}_n(\omega))/\alpha\bar{Y}_n(\omega) \to -Y(\omega)$. Since $-Y$ has

the distribution of N and $\overline{Y}_n$ has the distribution of $\overline{X}_n$, it follows that

$$\frac{\sqrt{n}}{\alpha}\left(\frac{1}{\overline{X}_n} - \alpha\right) \Rightarrow N;$$

thus $1/\overline{X}_n$ is approximately normally distributed with mean α and standard deviation $\alpha/\sqrt{n}$. In effect, $1/\overline{X}_n$ has been studied through the local linear approximation to the function $1/x$. This is called the *delta method*. ∎

The Lindeberg and Lyapounov Theorems

Suppose that for each n

(27.6) $X_{n1}, \ldots, X_{nr_n}$

are independent; the probability space for the sequence may change with n. Such a collection is called a *triangular array* of random variables. Put $S_n = X_{n1} + \cdots + X_{nr_n}$. Theorem 27.1 covers the special case in which $r_n = n$ and $X_{nk} = X_k$. Example 6.3 on the number of cycles in a random permutation shows that the idea of triangular array is natural and useful. The central limit theorem for triangular arrays will be applied in Example 27.3 to the same array.

To establish the asymptotic normality of S_n by means of the ideas in the preceding proof requires expanding the characteristic function of each X_{nk} to second-order terms and estimating the remainder. Suppose that the means are 0 and the variances are finite; write

(27.7) $E[X_{nk}] = 0,$ $\sigma_{nk}^2 = E[X_{nk}^2],$ $s_n^2 = \sum_{k=1}^{r_n} \sigma_{nk}^2.$

The assumption that X_{nk} has mean 0 entails no loss of generality. Assume $s_n^2 > 0$ for large n. A successful remainder estimate is possible under the assumption of the *Lindeberg condition*:

(27.8) $\lim_{n \to \infty} \sum_{k=1}^{r_n} \frac{1}{s_n^2} \int_{|X_{nk}| \geq \epsilon s_n} X_{nk}^2 \, dP = 0$

for $\epsilon > 0$.

Theorem 27.2. *Suppose that for each n the sequence $X_{n1}, \ldots, X_{nr_n}$ is independent and satisfies (27.7). If (27.8) holds for all positive ϵ, then $S_n/s_n \Rightarrow N$.*

This theorem contains the preceding one: Suppose that $X_{nk} = X_k$ and $r_n = n$, where the entire sequence $\{X_k\}$ is independent and the X_k all have the same distribution with mean 0 and variance σ^2. Then (27.8) reduces to

$$(27.9) \qquad \lim_{n \to \infty} \frac{1}{\sigma^2} \int_{|X_1| \ge \epsilon \sigma \sqrt{n}} X_1^2 \, dP = 0,$$

which holds because $[|X_1| \ge \epsilon \sigma \sqrt{n}] \downarrow \varnothing$ as $n \uparrow \infty$.

PROOF OF THE THEOREM.   Replacing X_{nk} by X_{nk}/s_n shows that there is no loss of generality in assuming

$$(27.10) \qquad s_n^2 = \sum_{k=1}^{r_n} \sigma_{nk}^2 = 1.$$

By (26.4_2),

$$\left| e^{itx} - \left(1 + itx - \tfrac{1}{2}t^2 x^2\right) \right| \le \min\{|tx|^2, |tx|^3\}.$$

Therefore, the characteristic function φ_{nk} of X_{nk} satisfies

$$(27.11) \qquad \left| \varphi_{nk}(t) - \left(1 - \tfrac{1}{2}t^2 \sigma_{nk}^2\right) \right| \le E\left[\min\{|tX_{nk}|^2, |tX_{nk}|^3\} \right].$$

Note that the expected value is finite.

For positive ϵ the right side of (27.11) is at most

$$\int_{|X_{nk}| < \epsilon} |tX_{nk}|^3 \, dP + \int_{|X_{nk}| \ge \epsilon} |tX_{nk}|^2 \, dP \le \epsilon |t|^3 \sigma_{nk}^2 + t^2 \int_{|X_{nk}| \ge \epsilon} X_{nk}^2 \, dP.$$

Since the σ_{nk}^2 add to 1 and ϵ is arbitrary, it follows by the Lindeberg condition that

$$(27.12) \qquad \sum_{k=1}^{r_n} \left| \varphi_{nk}(t) - \left(1 - \tfrac{1}{2}t^2 \sigma_{nk}^2\right) \right| \to 0$$

for each fixed t. The objective now is to show that

$$(27.13) \qquad \prod_{k=1}^{r_n} \varphi_{nk}(t) = \prod_{k=1}^{r_n} \left(1 - \tfrac{1}{2}t^2 \sigma_{nk}^2\right) + o(1)$$

$$= \prod_{k=1}^{r_n} e^{-t^2 \sigma_{nk}^2/2} + o(1) = e^{-t^2/2} + o(1).$$

For ϵ positive,

$$\sigma_{nk}^2 \le \epsilon^2 + \int_{|X_{nk}| \ge \epsilon} X_{nk}^2 \, dP,$$

and so it follows by the Lindeberg condition (recall that s_n is now 1) that

$$(27.14) \qquad\qquad \max_{1 \le k \le r_n} \sigma_{nk}^2 \to 0.$$

For large enough n, $1 - \frac{1}{2}t^2\sigma_{nk}^2$ are all between 0 and 1, and by (27.5), $\prod_{k=1}^{r_n}\varphi_{nk}(t)$ and $\prod_{k=1}^{r_n}(1 - \frac{1}{2}t^2\sigma_{nk}^2)$ differ by at most the sum in (27.12). This establishes the first of the asymptotic relations in (27.13).

Now (27.5) also implies that

$$\left| \prod_{k=1}^{r_n} e^{-t^2\sigma_{nk}^2/2} - \prod_{k=1}^{r_n}\left(1 - \tfrac{1}{2}t^2\sigma_{nk}^2\right) \right| \le \sum_{k=1}^{r_n} \left| e^{-t^2\sigma_{nk}^2/2} - 1 + \tfrac{1}{2}t^2\sigma_{nk}^2 \right|.$$

For complex z,

$$(27.15) \qquad\qquad |e^z - 1 - z| \le |z|^2 \sum_{k=2}^{\infty} \frac{|z|^{k-2}}{k!} \le |z|^2 e^{|z|}.$$

Using this in the right member of the preceding inequality bounds it by $t^4 e^{t^2}\sum_{k=1}^{r_n}\sigma_{nk}^4$; by (27.14) and (27.10), this sum goes to 0, from which the second equality in (27.13) follows. ∎

It is shown in the next section (Example 28.4) that if the independent array $\{X_{nk}\}$ satisfies (27.7), and if $S_n/s_n \Rightarrow N$, then the Lindeberg condition holds, *provided* $\max_{k \le r_n} \sigma_{nk}^2/s_n^2 \to 0$. But this converse fails without the extra condition: Take $X_{nk} = X_k$ normal with mean 0 and variance $\sigma_{nk}^2 = \sigma_k^2$, where $\sigma_1^2 = 1$ and $\sigma_n^2 = ns_{n-1}^2$.

Example 27.3. *Goncharov's theorem.* Consider the sum $S_n = \sum_{k=1}^{n} X_{nk}$ in Example 6.3. Here S_n is the number of cycles in a random permutation on n letters, the X_{nk} are independent, and

$$P[X_{nk} = 1] = \frac{1}{n - k + 1} = 1 - P[X_{nk} = 0].$$

The mean m_n is $L_n = \sum_{k=1}^{n} k^{-1}$, and the variance s_n^2 is $L_n + O(1)$. Lindeberg's condition for $X_{nk} - (n - k + 1)^{-1}$ is easily verified because these random variables are bounded by 1.

The theorem gives $(S_n - L_n)/s_n \Rightarrow N$. Now, in fact, $L_n = \log n + O(1)$, and so (see Example 25.8) the sum can be renormalized: $(S_n - \log n)/\sqrt{\log n} \Rightarrow N$. ∎

Suppose that the $|X_{nk}|^{2+\delta}$ are integrable for some positive δ and that *Lyapounov's condition*

$$(27.16) \qquad\qquad \lim_n \sum_{k=1}^{r_n} \frac{1}{s_n^{2+\delta}} E\left[|X_{nk}|^{2+\delta}\right] = 0$$

holds. Then Lindeberg's condition follows because the sum in (27.8) is bounded by

$$\sum_{k=1}^{r_n} \frac{1}{s_n^2} \int_{|X_{nk}| \geq \epsilon s_n} \frac{|X_{nk}|^{2+\delta}}{\epsilon^\delta s_n^\delta} \, dP \leq \frac{1}{\epsilon^\delta} \sum_{k=1}^{r_n} \frac{1}{s_n^{2+\delta}} E\left[|X_{nk}|^{2+\delta}\right].$$

Hence Theorem 27.2 has this corollary:

Theorem 27.3. *Suppose that for each n the sequence $X_{n1}, \ldots, X_{nr_n}$ is independent and satisfies (27.7). If (27.16) holds for some positive δ, then $S_n/s_n \Rightarrow N$.*

Example 27.4. Suppose that $X_1, X_2, \ldots$ are independent and uniformly bounded and have mean 0. If the variance s_n^2 of $S_n = X_1 + \cdots + X_n$ goes to ∞, then $S_n/s_n \Rightarrow N$: If K bounds the X_n, then

$$\sum_{k=1}^n \frac{1}{s_n^3} E\left[|X_k|^3\right] \leq \sum_{k=1}^n \frac{KE\left[X_k^2\right]}{s_n^3} = \frac{K}{s_n} \to 0,$$

which is Lyapounov's condition for $\delta = 1$.   ■

Example 27.5. Elements are drawn from a population of size n, randomly and with replacement, until the number of distinct elements that have been sampled is r_n, where $1 \leq r_n \leq n$. Let S_n be the drawing on which this first happens. A coupon collector requires S_n purchases to fill out a given portion of the complete set. Suppose that r_n varies with n in such a way that $r_n/n \to \rho$, $0 < \rho < 1$. What is the approximate distribution of S_n?

Let Y_p be the trial on which success first occurs in a Bernoulli sequence with probability p for success: $P[Y_p = k] = q^{k-1}p$, where $q = 1 - p$. Since the moment generating function is $pe^s/(1 - qe^s)$, $E[Y_p] = p^{-1}$ and $\text{Var}[Y_p] = qp^{-2}$. If $k - 1$ distinct items have thus far entered the sample, the waiting time until the next distinct one enters is distributed as Y_p as $p = (n - k + 1)/n$. Therefore, S_n can be represented as $\sum_{k=1}^{r_n} X_{nk}$ for independent summands X_{nk} distributed as $Y_{(n-k+1)/n}$. Since $r_n \sim \rho n$, the mean and variance above give

$$m_n = E[S_n] = \sum_{k=1}^{r_n} \left(1 - \frac{k-1}{n}\right)^{-1} \sim n \int_0^\rho \frac{dx}{1-x}$$

and

$$s_n^2 = \sum_{k=1}^{r_n} \frac{k-1}{n}\left(1 - \frac{k-1}{n}\right)^{-2} \sim n \int_0^p \frac{x\,dx}{(1-x)^2}.$$

Lyapounov's theorem applies for $\delta = 2$, and to check (27.16) requires the inequality

$$(27.17) \qquad E\left[\left(Y_p - p^{-1}\right)^4\right] \le Kp^{-4}$$

for some K independent of p. A calculation with the moment generating function shows that the left side is in fact $qp^{-4}(1 + 7q + q^2)$. It now follows that

$$(27.18) \qquad \sum_{k=1}^{r_n} E\left[\left(X_{nk} - \frac{n}{n-k+1}\right)^4\right] \le K \sum_{k=1}^{r_n}\left(1 - \frac{k-1}{n}\right)^{-4}$$

$$\sim Kn \int_0^p \frac{dx}{(1-x)^4}.$$

Since (27.16) follows from this, Theorem 27.3 applies: $(S_n - m_n)/s_n \Rightarrow N$. ∎

Dependent Variables*

The assumption of independence in the preceding theorems can be relaxed in various ways. Here a central limit theorem will be proved for sequences in which random variables far apart from one another are nearly independent in a sense to be defined.

For a sequence $X_1, X_2, \ldots$ of random variables, let α_n be a number such that

$$(27.19) \qquad |P(A \cap B) - P(A)P(B)| \le \alpha_n$$

for $A \in \sigma(X_1, \ldots, X_k)$, $B \in \sigma(X_{k+n}, X_{k+n+1}, \ldots)$, and $k \ge 1$, $n \ge 1$. Suppose that $\alpha_n \to 0$, the idea being that X_k and X_{k+n} are then approximately independent for large n. In this case the sequence $\{X_n\}$ is said to be α-mixing. If the distribution of the random vector $(X_n, X_{n+1}, \ldots, X_{n+j})$ does not depend on n, the sequence is said to be *stationary*.

Example 27.6. Let $\{Y_n\}$ be a Markov chain with finite state space and positive transition probabilities p_{ij}, and suppose that $X_n = f(Y_n)$, where f is some real function on the state space. If the initial probabilities p_i are the stationary ones (see Theorem 8.9), then clearly $\{X_n\}$ is stationary. Moreover, by (8.42), $|p_{ij}^{(n)} - p_j| \le \rho^n$, where $\rho < 1$. By (8.11), $P[Y_1 = i_1, \ldots, Y_k = i_k, Y_{k+n} = j_0, \ldots, Y_{k+n+l} = j_l] = p_{i_1} p_{i_1 i_2} \cdots p_{i_{k-1} i_k} p_{i_k j_0}^{(n)} p_{j_0 j_1} \cdots p_{j_{l-1} j_l}$, which differs

*This topic may be omitted.

from $P[Y_1 = i_1, \ldots, Y_k = i_k]P[Y_{k+n} = j_0, \ldots, Y_{k+n+l} = j_l]$ by at most $p_{i_1} p_{i_1 i_2} \cdots p_{i_{k-1} i_k} p^n p_{j_0 j_1} \cdots p_{j_{l-1} j_l}$. It follows by addition that, if s is the number of states, then for sets of the form $A = [(Y_1, \ldots, Y_k) \in H]$ and $B = [(Y_{k+n}, \ldots, Y_{k+n+l}) \in H']$, (27.19) holds with $\alpha_n = sp^n$. These sets (for k and n fixed) form fields generating σ-fields which contain $\sigma(X_1, \ldots, X_k)$ and $\sigma(X_{k+n}, X_{k+n+1}, \ldots)$, respectively. For fixed A the set of B satisfying (27.19) is a monotone class, and similarly if A and B are interchanged. It follows by the monotone class theorem (Theorem 3.4) that $\{X_n\}$ is α-mixing with $\alpha_n = sp^n$.   ∎

The sequence is *m-dependent* if $(X_1, \ldots, X_k)$ and $(X_{k+n}, \ldots, X_{k+n+l})$ are independent whenever $n > m$. In this case the sequence is α-mixing with $\alpha_n = 0$ for $n > m$. In this terminology an independent sequence is 0-dependent.

Example 27.7. Let $Y_1, Y_2, \ldots$ be independent and identically distributed, and put $X_n = f(Y_n, \ldots, Y_{n+m})$ for a real function f on R^{m+1}. Then $\{X_n\}$ is stationary and m-dependent.   ∎

Theorem 27.4. *Suppose that $X_1, X_2, \ldots$ is stationary and α-mixing with $\alpha_n = O(n^{-5})$ and that $E[X_n] = 0$ and $E[X_n^{12}] < \infty$. If $S_n = X_1 + \cdots + X_n$, then*

$$(27.20) \qquad n^{-1} \operatorname{Var}[S_n] \to \sigma^2 = E[X_1^2] + 2 \sum_{k=1}^{\infty} E[X_1 X_{1+k}],$$

where the series converges absolutely. If $\sigma > 0$, then $S_n / \sigma \sqrt{n} \Rightarrow N$.

The conditions $\alpha_n = O(n^{-5})$ and $E[X_n^{12}] < \infty$ are stronger than necessary; they are imposed to avoid technical complications in the proof. The idea of the proof, which goes back to Markov, is this: Split the sum $X_1 + \cdots + X_n$ into alternate blocks of length b_n (the big blocks) and l_n (the little blocks). Namely, let

$$(27.21) \quad U_{ni} = X_{(i-1)(b_n+l_n)+1} + \cdots + X_{(i-1)(b_n+l_n)+b_n}, \qquad 1 \le i \le r_n,$$

where r_n is the largest integer i for which $(i-1)(b_n + l_n) + b_n < n$. Further, let

$$(27.22) \quad \begin{aligned} V_{ni} &= X_{(i-1)(b_n+l_n)+b_n+1} + \cdots + X_{i(b_n+l_n)}, \qquad 1 \le i < r_n, \\ V_{nr_n} &= X_{(r_n-1)(b_n+l_n)+b_n+1} + \cdots + X_n. \end{aligned}$$

Then $S_n = \sum_{i=1}^{r_n} U_{ni} + \sum_{i=1}^{r_n} V_{ni}$, and the technique will be to choose the l_n small enough that $\sum_i V_{ni}$ is small in comparison with $\sum_i U_{ni}$ but large enough

that the U_{ni} are nearly independent, so that Lyapounov's theorem can be adapted to prove $\Sigma_i U_{ni}$ asymptotically normal.

Lemma 2. *If Y is measurable $\sigma(X_1, \ldots, X_k)$ and bounded by C, and if Z is measurable $\sigma(X_{k+n}, X_{k+n+1}, \ldots)$ and bounded by D, then*

$$(27.23) \qquad |E[YZ] - E[Y]E[Z]| \le 4CD\alpha_n.$$

PROOF. It is no restriction to take $C = D = 1$ and (by the usual approximation method) to take $Y = \Sigma_i y_i I_{A_i}$ and $Z = \Sigma_j z_j I_{B_j}$ simple ($|y_i|, |z_j| \le 1$). If $d_{ij} = P(A_i \cap B_j) - P(A_i)P(B_j)$, the left side of (27.23) is $|\Sigma_{ij} y_i z_j d_{ij}|$. Take ξ_i to be $+1$ or -1 as $\Sigma_j z_j d_{ij}$ is positive or not; now take η_j to be $+1$ or -1 as $\Sigma_i \xi_i d_{ij}$ is positive or not. Then

$$\left| \sum_{i,j} y_i z_j d_{ij} \right| \le \sum_i \left| \sum_j z_j d_{ij} \right| = \sum_i \xi_i \sum_j z_j d_{ij}$$

$$\le \sum_j \left| \sum_i \xi_i d_{ij} \right| = \sum_j \eta_j \sum_i \xi_i d_{ij} = \sum_{i,j} \xi_i \eta_j d_{ij}.$$

Let $A^{(0)} [B^{(0)}]$ be the union of the $A_i [B_j]$ for which $\xi_i = +1 [\eta_j = +1]$, and let $A^{(1)} = \Omega - A^{(0)} [B^{(1)} = \Omega - B^{(0)}]$. Then

$$\sum_{i,j} \xi_i \eta_j d_{ij} \le \sum_{u,\iota} |P(A^{(u)} \cap B^{(\iota)}) - P(A^{(u)})P(B^{(\iota)})| \le 4\alpha_n. \qquad \blacksquare$$

Lemma 3. *If Y is measurable $\sigma(X_1, \ldots, X_k)$ and $E[Y^4] \le C$, and if Z is measurable $\sigma(X_{k+n}, X_{k+n+1}, \ldots)$ and $E[Z^4] \le D$, then*

$$(27.24) \qquad |E[YZ] - E[Y]E[Z]| \le 8(1 + C + D)\alpha_n^{1/2}.$$

PROOF. Let $Y_0 = YI_{[|Y| \le a]}$, $Y_1 = YI_{[|Y| > a]}$, $Z_0 = ZI_{[|Z| \le a]}$, $Z_1 = ZI_{[|Z| > a]}$. By Lemma 2, $|E[Y_0 Z_0] - E[Y_0]E[Z_0]| \le 4a^2\alpha_n$. Further,

$$|E[Y_0 Z_1] - E[Y_0]E[Z_1]| \le E[|Y_0 - E[Y_0]| \cdot |Z_1 - E[Z_1]|]$$

$$\le 2a \cdot 2E[|Z_1|] \le 4aE[|Z_1| \cdot |Z_1/a|^3] \le 4D/a^2.$$

Similarly, $|E[Y_1 Z_0] - E[Y_1]E[Z_0]| \le 4C/a^2$. Finally,

$$|E[Y_1 Z_1] - E[Y_1]E[Z_1]| \le \mathrm{Var}^{1/2}[Y_1] \, \mathrm{Var}^{1/2}[Z_1] \le E^{1/2}[Y_1^2] E^{1/2}[Z_1^2]$$

$$\le E^{1/2}[Y_1^4/a^2] E^{1/2}[Z_1^4/a^2] \le C^{1/2}D^{1/2}/a^2.$$

Adding these inequalities gives $4a^2\alpha_n + 4(C + D)a^{-2} + C^{1/2}D^{1/2}a^{-2}$ as a

bound for the left side of (27.24). Take $a = \alpha_n^{-1/4}$ and observe that $4 + 4(C + D) + C^{1/2}D^{1/2} \leq 4 + 4(C^{1/2} + D^{1/2})^2 \leq 4 + 8(C + D)$. ∎

PROOF OF THEOREM 27.4. By Lemma 3, $|E[X_1 X_{1+n}]| \leq 8(1 + 2E[X_1^4])\alpha_n^{1/2} = O(n^{-5/2})$, and so the series in (27.20) converges absolutely. If $\rho_k = E[X_1 X_{1+k}]$, then by stationary $E[S_n^2] = n\rho_0 + 2\sum_{k=1}^{n-1}(n-k)\rho_k$ and therefore $|\sigma^2 - n^{-1}E[S_n^2]| \leq 2\sum_{k=n}^{\infty}|\rho_k| + 2n^{-1}\sum_{i=1}^{n-1}\sum_{k=i}^{\infty}|\rho_k|$; hence (27.20).

By stationarity again,

$$E[S_n^4] \leq 4! n \sum \left| E[X_1 X_{1+i} X_{1+i+j} X_{1+i+j+k}] \right|,$$

where the indices in the sum are constrained by $i, j, k \geq 0$ and $i + j + k < n$. By Lemma 3 the summand is at most

$$8\left(1 + E[X_1^4] + E[X_{1+i}^4 X_{1+i+j}^4 X_{1+i+j+k}^4]\right)\alpha_i^{1/2},$$

which is at most[†]

$$8\left(1 + E[X_1^4] + E[X_1^{12}]\right)\alpha_i^{1/2} = K_1 \alpha_i^{1/2}.$$

Similarly, $K_1 \alpha_k^{1/2}$ is a bound. Hence

$$E[S_n^4] \leq 4! n^2 \sum_{\substack{i, k \geq 0 \\ i+k < n}} K_1 \min\{\alpha_i^{1/2}, \alpha_k^{1/2}\}$$

$$\leq K_2 n^2 \sum_{0 \leq i \leq k} \alpha_k^{1/2} = K_2 n^2 \sum_{k=0}^{\infty} (k+1)\alpha_k^{1/2}.$$

Since $\alpha_k = O(k^{-5})$, the series here converges, and therefore

(27.25) $$E[S_n^4] \leq Kn^2$$

for some K independent of n.

Let $b_n = \lfloor n^{3/4} \rfloor$ and $l_n = \lfloor n^{1/4} \rfloor$. If r_n is the largest integer i such that $(i - 1)(b_n + l_n) + b_n < n$, then

(27.26) $$b_n \sim n^{3/4}, \qquad l_n \sim n^{1/4}, \qquad r_n \sim n^{1/4}.$$

Consider the random variables (27.21) and (27.22). By (27.25), (27.26), and

[†]$E|XYZ| \leq E^{1/3}|X|^3 \cdot E^{2/3}|YZ|^{3/2} \leq E^{1/3}|X|^3 \cdot E^{1/3}|Y|^3 \cdot E^{1/3}|Z|^3.$

stationarity,

$$P\left[\left|\frac{1}{\sigma\sqrt{n}}\sum_{i=1}^{r_n-1}V_{ni}\right|\ge\epsilon\right]\le\sum_{i=1}^{r_n-1}P\left[|V_{ni}|\ge\frac{\epsilon\sigma\sqrt{n}}{r_n}\right]$$

$$\le\frac{r_n^4}{\epsilon^4\sigma^4 n^2}r_n Kl_n^2\sim\frac{K}{\epsilon^4\sigma^4 n^{1/4}}\to 0;$$

(27.25) and (27.26) also give

$$P\left[\frac{1}{\sigma\sqrt{n}}|V_{nr_n}|\ge\epsilon\right]\le\frac{K(b_n+l_n)^2}{\epsilon^4\sigma^4 n^2}\sim\frac{K}{\epsilon^4\sigma^4 n^{1/2}}\to 0.$$

Therefore, $\sum_{i=1}^{r_n}V_{ni}/\sigma\sqrt{n}\Rightarrow 0$, and by Theorem 25.4 it suffices to prove that $\sum_{i=1}^{r_n}U_{ni}/\sigma\sqrt{n}\Rightarrow N$.

Let U_{ni}', $1\le i\le r_n$, be independent random variables having the distribution common to the U_{ni}. By Lemma 2 extended inductively the characteristic functions of $\sum_{i=1}^{r_n}U_{ni}/\sigma\sqrt{n}$ and of $\sum_{i=1}^{r_n}U_{ni}'/\sigma\sqrt{n}$ differ by at most[†] $16r_n\alpha_{l_n}$. Since $\alpha_n=O(n^{-5})$, this difference is $O(n^{-1})$ by (27.26).

The characteristic function of $\sum_{i=1}^{r_n}U_{ni}/\sigma\sqrt{n}$ will thus approach $e^{-t^2/2}$ if that of $\sum_{i=1}^{r_n}U_{ni}'/\sigma\sqrt{n}$ does. It therefore remains only to show that $\sum_{i=1}^{r_n}U_{ni}'/\sigma\sqrt{n}\Rightarrow N$. Now $E[|U_{ni}'|^2]=E[U_{n1}^2]\sim b_n\sigma^2$ by (27.20). Further, $E[|U_{ni}'|^4]\le Kb_n^2$ by (27.25). Lyapounov's condition (27.16) for $\delta=2$ therefore follows because

$$\frac{r_n E\left[|U_{n1}'|^4\right]}{\left(r_n E\left[|U_{n1}'|^2\right]\right)^2}\sim\frac{E\left[|U_{n1}'|^4\right]}{r_n b_n^2\sigma^4}\le\frac{K}{r_n\sigma^4}\to 0.\qquad\blacksquare$$

Example 27.8. Let $\{Y_n\}$ be the stationary Markov process of Example 27.6. Let f be a function on the state space, put $m=\sum_i p_i f(i)$, and define $X_n=f(Y_n)-m$. Then $\{X_n\}$ satisfies the conditions of Theorem 27.4. If $\beta_{ij}=\delta_{ij}p_i-p_i p_j+2p_i\sum_{k=1}^\infty(p_{ij}^{(k)}-p_j)$, then the σ^2 in (27.20) is $\sum_{ij}\beta_{ij}(f(i)-m)(f(j)-m)$, and $\sum_{k=1}^n f(Y_k)$ is approximately normally distributed with mean nm and standard deviation $\sigma\sqrt{n}$.

If $f(i)=\delta_{i_0 i}$, then $\sum_{k=1}^n f(Y_k)$ is the number of passages through the state i_0 in the first n steps of the process. In this case $m=p_{i_0}$ and $\sigma^2=p_{i_0}(1-p_{i_0})+2p_{i_0}\sum_{k=1}^\infty(p_{i_0 i_0}^{(k)}-p_{i_0})$. $\qquad\blacksquare$

Example 27.9. If the X_n are stationary and m-dependent and have mean 0, Theorem 27.4 applies and $\sigma^2=E[X_1^2]+2\sum_{k=1}^m E[X_1 X_{1+k}]$. Example 27.7 is a case in point. Taking $m=1$ and $f(x,y)=x-y$ in that example gives an instance where $\sigma^2=0$. $\qquad\blacksquare$

[†] The 4 in (27.23) has become 16 to allow for splitting into real and imaginary parts.

PROBLEMS

27.1. Prove Theorem 23.2 by means of characteristic functions. *Hint*: Use (27.5) to compare the characteristic function of $\sum_{k=1}^{r_n} Z_{nk}$ with $\exp[\sum_k p_{nk}(e^{it} - 1)]$.

27.2. If $\{X_n\}$ is independent and the X_n all have the same distribution with finite first moment, then $n^{-1}S_n \to E[X_1]$ with probability 1 (Theorem 22.1), so that $n^{-1}S_n \Rightarrow E[X_1]$. Prove the latter fact by characteristic functions. *Hint*: Use (27.5).

27.3. For a Poisson variable Y_λ with mean λ, show that $(Y_\lambda - \lambda)/\sqrt{\lambda} \Rightarrow N$ as $\lambda \to \infty$. Show that (22.3) fails for $t = 1$.

27.4. Suppose that $|X_{nk}| \le M_n$ with probability 1 and $M_n/s_n \to 0$. Verify Lyapounov's condition and then Lindeberg's condition.

27.5. Suppose that the random variables in any single row of the triangular array are identically distributed. To what do Lindeberg's and Lyapounov's conditions reduce?

27.6. Suppose that $Z_1, Z_2, \ldots$ are independent and identically distributed with mean 0 and variance 1, and suppose that $X_{nk} = \sigma_{nk} Z_k$. Write down the Lindeberg condition and show that it holds if $\max_{k \le r_n} \sigma_{nk}^2 = o(\sum_{k=1}^{r_n} \sigma_{nk}^2)$.

27.7. Construct an example where Lindeberg's condition holds but Lyapounov's does not.

27.8. 22.9↑   Prove a central limit theorem for the number R_n of records up to time n.

27.9. 6.3↑   Let S_n be the number of inversions in a random permutation on n letters. Prove a central limit theorem for S_n.

27.10. *The δ-method.* Suppose that Theorem 27.1 applies to $\{X_n\}$, so that $\sqrt{n}\,\sigma^{-1}(\overline{X}_n - c) \Rightarrow N$, where $\overline{X}_n = n^{-1}\sum_{k=1}^n X_k$. Use Theorem 25.6 as in Example 27.2 to show that, if $f(x)$ has a nonzero derivative at c, then $\sqrt{n}\,(f(\overline{X}_n) - f(c))/\sigma|f'(c)| \Rightarrow N$: $\overline{X}_n$ is approximately normal with mean c and standard deviation $\sigma/\sqrt{n}$, and $f(\overline{X}_n)$ is approximately normal with mean $f(c)$ and standard deviation $|f'(c)|\sigma/\sqrt{n}$. Example 27.2 is the case $f(x) = 1/x$.

27.11. Suppose independent X_n have density $|x|^{-3}$ outside $(-1, +1)$. Show that $(n \log n)^{-1/2} S_n \Rightarrow N$.

27.12. There can be asymptotic normality even if there are no moments at all. Construct a simple example.

27.13. Let $d_n(\omega)$ be the dyadic digits of a point ω drawn at random from the unit interval. For a k-tuple $(u_1, \ldots, u_k)$ of 0's and 1's, let $N_n(u_1, \ldots, u_k; \omega)$ be the number of $m \le n$ for which $(d_m(\omega), \ldots, d_{m+k-1}(\omega)) = (u_1, \ldots, u_k)$. Prove a central limit theorem for $N_n(u_1, \ldots, u_k; \omega)$. (See Problem 6.12.)

27.14. *The central limit theorem for a random number of summands.* Let $X_1, X_2, \ldots$ be independent, identically distributed random variables with mean 0 and variance σ^2, and let $S_n = X_1 + \cdots + X_n$. For each positive t, let ν_t be a random variable assuming positive integers as values; it need not be independent of the X_n. Suppose that there exist positive constants a_t and θ such that

$$a_t \to \infty, \qquad \frac{\nu_t}{a_t} \Rightarrow \theta$$

as $t \to \infty$. Show by the following steps that

(27.27) $$\frac{S_{\nu_t}}{\sigma\sqrt{\nu_t}} \Rightarrow N, \qquad \frac{S_{\nu_t}}{\sigma\sqrt{\theta a_t}} \Rightarrow N.$$

(a) Show that it may be assumed that $\theta = 1$ and the a_t are integers.
(b) Show that it suffices to prove the second relation in (27.27).
(c) Show that it suffices to prove $(S_{\nu_t} - S_{a_t})/\sqrt{a_t} \Rightarrow 0$.
(d) Show that

$$P\left[|S_{\nu_t} - S_{a_t}| \ge \epsilon\sqrt{a_t}\right] \le P\left[|\nu_t - a_t| \ge \epsilon^3 a_t\right]$$

$$+ P\left[\max_{|k - a_t| \le \epsilon^3 a_t} |S_k - S_{a_t}| \ge \epsilon\sqrt{a_t}\right],$$

and conclude from Kolmogorov's inequality that the last probability is at most $2\epsilon\sigma^2$

27.15. 21.21 23.10 23.14↑ *A central limit theorem in renewal theory.* Let $X_1, X_2, \ldots$ be independent, identically distributed positive random variables with mean m and variance σ^2, and as in Problem 23.10 let N_t be the maximum n for which $S_n \le t$. Prove by the following steps that

$$\frac{N_t - tm^{-1}}{\sigma t^{1/2} m^{-3/2}} \Rightarrow N.$$

(a) Show by the results in Problems 21.21 and 23.10 that $(S_{N_t} - t)/\sqrt{t} \Rightarrow 0$.
(b) Show that it suffices to prove that

$$\frac{N_t - S_{N_t} m^{-1}}{\sigma t^{1/2} m^{-3/2}} = \frac{-(S_{N_t} - mN_t)}{\sigma t^{1/2} m^{-1/2}} \Rightarrow N.$$

(c) Show (Problem 23.10) that $N_t/t \Rightarrow m^{-1}$, and apply the theorem in Problem 27.14.

27.16. Show by partial integration that

$$(27.28) \qquad \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-u^2/2} \, du \sim \frac{1}{\sqrt{2\pi}} \frac{1}{x} e^{-x^2/2}$$

as $x \to \infty$.

27.17. ↑ Suppose that $X_1, X_2, \ldots$ are independent and identically distributed with mean 0 and variance 1, and suppose that $a_n \to \infty$. Formally combine the central limit theorem and (27.28) to obtain

$$(27.29) \qquad P\left[S_n \ge a_n \sqrt{n} \right] \sim \frac{1}{\sqrt{2\pi}} \frac{1}{a_n} e^{-a_n^2/2} = e^{-a_n^2(1+\zeta_n)/2},$$

where $\zeta_n \to 0$ if $a_n \to \infty$. For a case in which this does hold, see Theorem 9.4.

27.18. 21.2↑ *Stirling's formula.* Let $S_n = X_1 + \cdots + X_n$, where the X_n are independent and each has the Poisson distribution with parameter 1. Prove successively.

(a) $E\left[\left(\frac{S_n - n}{\sqrt{n}} \right)^- \right] = e^{-n} \sum_{k=0}^n \left(\frac{n-k}{\sqrt{n}} \right) \frac{n^k}{k!} = \frac{n^{n+(1/2)} e^{-n}}{n!}.$

(b) $\left(\frac{S_n - n}{\sqrt{n}} \right)^- \Rightarrow N^-.$

(c) $E\left[\left(\frac{S_n - n}{\sqrt{n}} \right)^- \right] \to E[N^-] = \frac{1}{\sqrt{2\pi}}.$

(d) $n! \sim \sqrt{2\pi} n^{n+(1/2)} e^{-n}.$

27.19. Let $l_n(\omega)$ be the length of the run of 0's starting at the nth place in the dyadic expansion of a point ω drawn at random from the unit interval; see Example 4.1.

(a) Show that $l_1, l_2, \ldots$ is an α-mixing sequence, where $\alpha_n = 4/2^n$.

(b) Show that $\sum_{k=1}^n l_k$ is approximately normally distributed with mean n and variance $6n$.

27.20. Prove under the hypotheses of Theorem 27.4 that $S_n/n \to 0$ with probability 1. *Hint*: Use (27.25).

27.21. 26.1 26.29↑ Let $X_1, X_2, \ldots$ be independent and identically distributed, and suppose that the distribution common to the X_n is supported by $[0, 2\pi]$ and is not a lattice distribution. Let $S_n = X_1 + \cdots + X_n$, where the sum is reduced modulo 2π. Show that $S_n \Rightarrow U$, where U is uniformly distributed over $[0, 2\pi]$.

SECTION 28. INFINITELY DIVISIBLE DISTRIBUTIONS*

Suppose that Z_λ has the Poisson distribution with parameter λ and that $X_{n1}, \ldots, X_{nn}$ are independent and $P[X_{nk} = 1] = \lambda/n$, $P[X_{nk} = 0] = 1 - \lambda/n$. According to Example 25.2, $X_{n1} + \cdots + X_{nn} \Rightarrow Z_\lambda$. This contrasts with the central limit theorem, in which the limit law is normal. What is the class of all possible limit laws for independent triangular arrays? A suitably restricted form of this question will be answered here.

Vague Convergence

The theory requires two preliminary facts about convergence of measures Let μ_n and μ be finite measures on $(R^1, \mathscr{R}^1)$. If $\mu_n(a, b] \to \mu(a, b]$ for every finite interval for which $\mu\{a\} = \mu\{b\} = 0$, then μ_n *converges vaguely* to μ, written $\mu_n \to_v \mu$. If μ_n and μ are probability measures, it is not hard to see that this is equivalent to weak convergence $\mu_n \Rightarrow \mu$. On the other hand, if μ_n is a unit mass at n and $\mu(R^1) = 0$, then $\mu_n \to_v \mu$, but $\mu_n \Rightarrow \mu$ makes no sense, because μ is not a probability measure.

The first fact needed is this: *Suppose that* $\mu_n \to_v \mu$ *and*

$$(28.1) \qquad\qquad \sup_n \mu_n(R^1) < \infty;$$

then

$$(28.2) \qquad\qquad \int f \, d\mu_n \to \int f \, d\mu$$

for every continuous real f that vanishes at $\pm\infty$ in the sense that $\lim_{|x| \to \infty} f(x) = 0$. Indeed, choose M so that $\mu(R^1) < M$ and $\mu_n(R^1) < M$ for all n. Given ϵ, choose a and b so that $\mu\{a\} = \mu\{b\} = 0$ and $|f(x)| < \epsilon/M$ if $x \notin A = (a, b]$. Then $|\int_{A^c} f \, d\mu_n| < \epsilon$ and $|\int_{A^c} f \, d\mu| < \epsilon$. If $\mu(A) > 0$, define $\nu(B) = \mu(B \cap A)/\mu(A)$ and $\nu_n(B) = \mu_n(B \cap A)/\mu_n(A)$. It is easy to see that $\nu_n \Rightarrow \nu$, so that $\int f \, d\nu_n \to \int f \, d\nu$. But then $|\int_A f \, d\mu_n - \int_A f \, d\mu| < \epsilon$ for large n, and hence $|\int f \, d\mu_n - \int f \, d\mu| < 3\epsilon$ for large n. If $\mu(A) = 0$, then $\int_A f \, d\mu_n \to 0$, and the argument is even simpler.

The other fact needed below is this: *If* (28.1) *holds, then there is a subsequence* $\{\mu_{n_k}\}$ *and a finite measure μ such that* $\mu_{n_k} \to_v \mu$ *as* $k \to \infty$. Indeed, let $F_n(x) = \mu_n(-\infty, x]$. Since the F_n are uniformly bounded because of (28.1), the proof of Helly's theorem shows there exists a subsequence $\{F_{n_k}\}$ and a bounded, nondecreasing, right-continuous function F such that $\lim_k F_{n_k}(x) = F(x)$ at continuity points x of F. If μ is the measure for which $\mu(a, b] = F(b) - F(a)$ (Theorem 12.4), then clearly $\mu_{n_k} \to_v \mu$.

The Possible Limits

Let $X_{n1}, \ldots, X_{nr_n}$, $n = 1, 2, \ldots$, be a triangular array as in the preceding section. The random variables in each row are independent, the means are 0,

and the variances are finite:

$$(28.3) \qquad E[X_{nk}] = 0, \qquad \sigma_{nk}^2 = E[X_{nk}^2], \qquad s_n^2 = \sum_{k=1}^{r_n} \sigma_{nk}^2.$$

Assume $s_n^2 > 0$ and put $S_n = X_{n1} + \cdots + X_{nr_n}$. Here it will be assumed that the total variance is bounded:

$$(28.4) \qquad \sup_n s_n^2 < \infty.$$

In order that the X_{nk} be small compared with S_n, assume that

$$(28.5) \qquad \lim_n \max_{k \le r_n} \sigma_{nk}^2 = 0.$$

The arrays in the preceding section were normalized by replacing X_{nk} by X_{nk}/s_n. This has the effect of replacing s_n by 1, in which case of course (28.4) holds, and (28.5) is the same thing as $\max_k \sigma_{nk}^2/s_n^2 \to 0$.

A distribution function F is *infinitely divisible* if for each n there is a distribution function F_n such that F is the n-fold convolution $F_n * \cdots * F_n$ (n copies) of F_n. The class of possible limit laws will turn out to consist of the infinitely divisible distributions with mean 0 and finite variance.[†] It will be possible to exhibit the characteristic functions of these laws in an explicit way.

Theorem 28.1. *Suppose that*

$$(28.6) \qquad \varphi(t) = \exp \int_{R^1} \left(e^{itx} - 1 - itx \right) \frac{1}{x^2} \mu(dx),$$

where μ is a finite measure. Then φ is the characteristic function of an infinitely divisible distribution with mean 0 and variance $\mu(R^1)$.

By (26.4$_2$), the integrand in (28.6) converges to $-t^2/2$ as $x \to 0$; take this as its value at $x = 0$. By (26.4$_1$), the integrand is at most $t^2/2$ in modulus and so is integrable.

The formula (28.6) is the *canonical representation* of φ, and μ is the *canonical measure*.

Before proceeding to the proof, consider three examples.

Example 28.1. If μ consists of a mass of σ^2 at the origin, (28.6) is $e^{-\sigma^2 t^2/2}$, the characteristic function of a centered normal distribution F. It is certainly infinitely divisible—take F_n normal with variance σ^2/n. ■

[†]There do exist infinitely divisible distributions without moments (see Problems 28.3 and 28.4), but they do not figure in the theory of this section

Example 28.2. Suppose that μ consists of a mass of λx^2 at $x \neq 0$. Then (28.6) is $\exp \lambda (e^{itx} - 1 - itx)$; but this is the characteristic function of $x(Z_\lambda - \lambda)$, where Z_λ has the Poisson distribution with mean λ. Thus (28.6) is the characteristic function of a distribution function F, and F is infinitely divisible—take F_n to be the distribution function of $x(Z_{\lambda/n} - \lambda/n)$. ∎

Example 28.3. If $\varphi_j(t)$ is given by (28.6) with μ_j for the measure, and if $\mu = \sum_{j=1}^k \mu_j$, then (28.6) is $\varphi_1(t)\ldots\varphi_k(t)$. It follows by the preceding two examples that (28.6) is a characteristic function if μ consists of finitely many point masses. It is easy to check in the preceding two examples that the distribution corresponding to $\varphi(t)$ has mean 0 and variance $\mu(R^1)$, and since the means and variances add, the same must be true in the present example. ∎

PROOF OF THEOREM 28.1. Let μ_k have mass $\mu(j2^{-k}, (j+1)2^{-k}]$ at $j2^{-k}$ for $j = 0, \pm 1, \ldots, \pm 2^{2k}$. Then $\mu_k \to_t \mu$. As observed in Example 28.3, if $\varphi_k(t)$ is (28.6) with μ_k in place of μ, then φ_k is a characteristic function. For each t the integrand in (28.6) vanishes at $\pm\infty$; since $\sup_k \mu_k(R^1) < \infty$, $\varphi_k(t) \to \varphi(t)$ follows (see (28.2)). By Corollary 2 to Theorem 26.3, $\varphi(t)$ is itself a characteristic function. Further, the distribution corresponding to $\varphi_k(t)$ has second moment $\mu_k(R^1)$, and since this is bounded, it follows (Theorem 25.11) that the distribution corresponding to $\varphi(t)$ has a finite second moment. Differentiation (use Theorem 16.8) shows that the mean is $\varphi'(0) = 0$ and the variance is $-\varphi''(0) = \mu(R^1)$. Thus (28.6) is always the characteristic function of a distribution with mean 0 and variance $\mu(R^1)$.

If $\psi_n(t)$ is (28.6) with μ/n in place of μ, then $\varphi(t) = \psi_n^n(t)$, so that the distribution corresponding to $\varphi(t)$ is indeed infinitely divisible. ∎

The representation (28.6) shows that the normal and Poisson distributions are special cases in a very large class of infinitely divisible laws.

Theorem 28.2. *Every infinitely divisible distribution with mean 0 and finite variance is the limit law of S_n for some independent triangular array satisfying* (28.3), (28.4), *and* (28.5). ∎

The proof requires this preliminary result:

Lemma. *If X and Y are independent and $X + Y$ has a second moment, then X and Y have second moments as well.*

PROOF. Since $X^2 + Y^2 \leq (X + Y)^2 + 2|XY|$, it suffices to prove $|XY|$ integrable, and by Fubini's theorem applied to the joint distribution of X and Y it suffices to prove $|X|$ and $|Y|$ individually integrable. Since $|Y| \leq |x| + |x + Y|$, $E[|Y|] = \infty$ would imply $E[|x + Y|] = \infty$ for each x; by Fubini's

theorem again $E[|Y|] = \infty$ would therefore imply $E[|X + Y|] = \infty$, which is impossible. Hence $E[|Y|] < \infty$, and similarly $E[|X|] < \infty$.   ∎

PROOF OF THEOREM 28.2. Let F be infinitely divisible with mean 0 and variance σ^2. If F is the n-fold convolution of F_n, then by the lemma (extended inductively) F_n has finite mean and variance, and these must be 0 and σ^2/n. Take $r_n = n$ and take $X_{n1}, \dots, X_{nn}$ independent, each with distribution function F_n.   ∎

Theorem 28.3. *If F is the limit law of S_n for an independent triangular array satisfying (28.3), (28.4), and (28.5), then F has characteristic function of the form (28.6) for some finite measure μ.*

PROOF. The proof will yield information making it possible to identify the limit. Let $\varphi_{nk}(t)$ be the characteristic function of X_{nk}. The first step is to prove that

$$(28.7) \qquad \prod_{k=1}^{r_n} \varphi_{nk}(t) - \exp \sum_{k=1}^{r_n} \left(\varphi_{nk}(t) - 1 \right) \to 0$$

for each t. Since $|z| \le 1$ implies that $|e^{z-1}| = e^{\mathrm{Re}\, z - 1} \le 1$, it follows by (27.5) that the difference $\delta_n(t)$ in (28.7) satisfies $|\delta_n(t)| \le \sum_{k=1}^{r_n} |\varphi_{nk}(t) - \exp(\varphi_{nk}(t) - 1)|$. Fix t. If $\varphi_{nk}(t) - 1 = \theta_{nk}$, then $|\theta_{nk}| \le t^2 \sigma_{nk}^2/2$, and it follows by (28.4) and (28.5) that $\max_k |\theta_{nk}| \to 0$ and $\sum_k |\theta_{nk}| = O(1)$. Therefore, for sufficiently large n, $|\delta_n(t)| \le \sum_k |1 + \theta_{nk} - e^{\theta_{nk}}| \le e^2 \sum_k |\theta_{nk}|^2 \le e^2 \max_k |\theta_{nk}| \cdot \sum_k |\theta_{nk}|$ by (27.15). Hence (28.7).

If F_{nk} is the distribution function of X_{nk}, then

$$\sum_{k=1}^{r_n} \left(\varphi_{nk}(t) - 1 \right) = \sum_{k=1}^{r_n} \int_{R^1} \left(e^{itx} - 1 \right) dF_{nk}(x)$$

$$= \sum_{k=1}^{r_n} \int_{R^1} \left(e^{itx} - 1 - itx \right) dF_{nk}(x).$$

Let μ_n be the finite measure satisfying

$$(28.8) \qquad \mu_n(-\infty, x] = \sum_{k=1}^{r_n} \int_{y \le x} y^2 \, dF_{nk}(y),$$

and put

$$(28.9) \qquad \varphi_n(t) = \exp \int_{R^1} \left(e^{itx} - 1 - itx \right) \frac{1}{x^2} \mu_n(dx).$$

Then (28.7) can be written

$$(28.10) \qquad \prod_{k=1}^{r_n} \varphi_{nk}(t) - \varphi_n(t) \to 0.$$

By (28.8), $\mu_n(R^1) = s_n^2$, and this is bounded by assumption. Thus (28.1) holds, and some subsequence $\{\mu_{n_u}\}$ converges vaguely to a finite measure μ. Since the integrand in (28.9) vanishes at $\pm\infty$, $\varphi_{n_u}(t)$ converges to (28.6). But, of course, $\lim_n \varphi_n(t)$ must coincide with the characteristic function of the limit law F, which exists by hypothesis. Thus F must have characteristic function of the form (28.6).   ∎

Theorems 28.1, 28.2, and 28.3 together show that the possible limit laws are exactly the infinitely divisible distributions with mean 0 and finite variance, and they give explicitly the form the characteristic functions of such laws must have.

Characterizing the Limit

Theorem 28.4. *Suppose that F has characteristic function (28.6) and that an independent triangular array satisfies (28.3), (28.4), and (28.5). Then S_n has limit law F if and only if $\mu_n \to_v \mu$, where μ_n is defined by (28.8).*

PROOF. Since (28.7) holds as before, S_n has limit law F if and only if $\varphi_n(t)$ (defined by (28.9)) converges for each t to $\varphi(t)$ (defined by (28.6)). If $\mu_n \to_v \mu$, then $\varphi_n(t) \to \varphi(t)$ follows because the integrand in (28.9) and (28.6) vanishes at $\pm\infty$ and because (28.1) follows from (28.4).

Now suppose that $\varphi_n(t) \to \varphi(t)$. Since $\mu_n(R^1) = s_n^2$ is bounded, each subsequence $\{\mu_{n_u}\}$ contains a further subsequence $\{\mu_{n_{u(i)}}\}$ converging vaguely to some ν. If it can be shown that ν necessarily coincides with μ, it will follow by the usual argument that $\mu_n \to_v \mu$. But by the definition (28.9) of $\varphi_n(t)$, it follows that $\varphi(t)$ must coincide with $\psi(t) = \exp \int_{R^1}(e^{itx} - 1 - itx)x^{-2}\nu(dx)$. Now $\varphi'(t) = i\varphi(t)\int_{R^1}(e^{itx} - 1)x^{-1}\mu(dx)$, and similarly for $\psi'(t)$. Hence $\varphi(t) = \psi(t)$ implies that $\int_{R^1}(e^{itx} - 1)x^{-1}\nu(dx) = \int_{R^1}(e^{itx} - 1)x^{-1}\mu(dx)$. A further differentiation gives $\int_R e^{itx}\mu(dx) = \int_R e^{itx}\nu(dx)$. This implies that $\mu(R^1) = \nu(R^1)$, and so $\mu = \nu$ by the uniqueness theorem for characteristic functions.   ∎

Example 28.4. *The normal case.* According to the theorem, $S_n \Rightarrow N$ if and only if μ_n converges vaguely to a unit mass at 0. If $s_n^2 = 1$, this holds if and only if $\sum_{k=1}^{r_n} \int_{|x| \geq \epsilon} x^2 \, dF_{nk}(x) \to 0$, which is exactly Lindeberg's condition.   ∎

Example 28.5. *The Poisson case.* Let $Z_{n1}, \ldots, Z_{nr_n}$ be an independent triangular array, and suppose $X_{nk} = Z_{nk} - m_{nk}$ satisfies the conditions of the

theorem, where $m_{nk} = E[Z_{nk}]$. If Z_λ has the Poisson distribution with parameter λ, then $\Sigma_k X_{nk} \Rightarrow Z_\lambda - \lambda$ if and only if μ_n converges vaguely to a mass of λ at 1 (see Example 28.2). If $s_n^2 \to \lambda$, the requirement is $\mu_n[1 - \epsilon, 1 + \epsilon] \to \lambda$, or

$$(28.11) \qquad \sum_k \int_{|Z_{nk} - m_{nk} - 1| > \epsilon} (Z_{nk} - m_{nk})^2 \, dP \to 0$$

for positive ϵ. If s_n^2 and $\Sigma_k m_{nk}$ both converge to λ, (28.11) is a necessary and sufficient condition for $\Sigma_k Z_{nk} \Rightarrow Z_\lambda$. The conditions are easily checked under the hypotheses of Theorem 23.2: Z_{nk} assumes the values 1 and 0 with probabilities p_{nk} and $1 - p_{nk}$, $\Sigma_k p_{nk} \to \lambda$, and $\max_k p_{nk} \to 0$. ∎

PROBLEMS

28.1. Show that $\mu_n \to_t \mu$ implies $\mu(R^1) \le \liminf_n \mu_n(R^1)$. Thus in vague convergence mass can "escape to infinity" but mass cannot "enter from infinity."

28.2. (a) Show that $\mu_n \to_t \mu$ if and only if (28.2) holds for every continuous f with bounded support.

(b) Show that if $\mu_n \to_t \mu$ but (28.1) does not hold, then there is a continuous f vanishing at $\pm\infty$ for which (28.2) does not hold.

28.3. 23.7↑ Suppose that $N, Y_1, Y_2, \ldots$ are independent, the Y_n have a common distribution function F, and N has the Poisson distribution with mean α. Then $S = Y_1 + \cdots + Y_N$ has the *compound Poisson distribution*.

(a) Show that the distribution of S is infinitely divisible. Note that S may not have a mean.

(b) The distribution function of S is $\Sigma_{n=0}^\infty e^{-\alpha} \alpha^n F^{n*}(x)/n!$, where F^{n*} is the n-fold convolution of F (a unit jump at 0 for $n = 0$). The characteristic function of S is $\exp \alpha \int_{-\infty}^\infty (e^{itx} - 1) \, dF(x)$.

(c) Show that, if F has mean 0 and finite variance, then the canonical measure μ in (28.6) is specified by $\mu(A) = \alpha \int_A x^2 \, dF(x)$.

28.4. (a) Let ν be a finite measure, and define

$$(28.12) \qquad \varphi(t) = \exp\left[i\gamma t + \int_{-\infty}^\infty \left(e^{itx} - 1 - \frac{itx}{1 + x^2} \right) \frac{1 + x^2}{x^2} \nu(dx) \right],$$

where the integrand is $-t^2/2$ at the origin. Show that this is the characteristic function of an infinitely divisible distribution.

(b) Show that the Cauchy distribution (see the table on p. 348) is the case where $\gamma = 0$ and ν has density $\pi^{-1}(1 + x^2)^{-1}$ with respect to Lebesgue measure.

28.5. Show that the Cauchy, exponential, and gamma (see (20.47)) distributions are infinitely divisible.

28.6. Find the canonical representation (28.6) of the exponential distribution with mean 1:

(a) The characteristic function is $\int_0^\infty e^{itx} e^{-x}\, dx = (1 - it)^{-1} = \varphi(t)$.

(b) Show that (use the principal branch of the logarithm or else operate formally for the moment) $d(\log \varphi(t))/dt = i\varphi(t) = i\int_0^\infty e^{itx} e^{-x}\, dx$. Integrate with respect to t to obtain

(28.13) $$\frac{1}{1 - it} = \exp \int_0^\infty \left(e^{itx} - 1\right) \frac{e^{-x}}{x}\, dx$$

Verify (28 13) after the fact by showing that the ratio of the two sides has derivative 0.

(c) Multiply (28.13) by e^{-it} to center the exponential distribution at its mean: The canonical measure μ has density xe^{-x} over $(0, \infty)$.

28.7. $\uparrow$ If X and Y are independent and each has the exponential density e^{-x}, then $X - Y$ has the double exponential density $\frac{1}{2} e^{-|x|}$ (see the table on p. 348). Show that its characteristic function is

$$\frac{1}{1 + t^2} = \exp \int_{-\infty}^\infty \left(e^{itx} - 1 - itx\right) \frac{1}{x^2} |x| e^{-|x|}\, dx.$$

28.8. $\uparrow$ Suppose $X_1, X_2, \ldots$ are independent and each has the double exponential density. Show that $\sum_{n=1}^\infty X_n/n$ converges with probability 1. Show that the distribution of the sum is infinitely divisible and that its canonical measure has density $|x| e^{-|x|}/(1 - e^{-|x|}) = \sum_{n=1}^\infty |x| e^{-|nx|}$.

28.9. 26.8$\uparrow$ Show that for the gamma density $e^{-x} x^{u-1}/\Gamma(u)$ the canonical measure has density uxe^{-x} over $(0, \infty)$.

The remaining problems require the notion of a *stable law*. A distribution function F is stable if for each n there exist constants a_n and b_n, $a_n > 0$, such that, if $X_1, \ldots, X_n$ are independent and have distribution function F, then $a_n^{-1}(X_1 + \cdots + X_n) + b_n$ also has distribution function F.

28.10. Suppose that for all a, a', b, b' there exist a'', b'' (here a, a', a'' are all positive) such that $F(ax + b) * F(a'x + b') = F(a''x + b'')$. Show that F is stable.

28.11. Show that a stable law is infinitely divisible.

28.12. Show that the Poisson law, although infinitely divisible, is not stable.

28.13. Show that the normal and Cauchy laws are stable.

28.14. 28.10$\uparrow$ Suppose that F has mean 0 and variance 1 and that the dependence of a'', b'' on a, a', b, b' is such that

$$F\left(\frac{x}{\sigma_1}\right) * F\left(\frac{x}{\sigma_2}\right) = F\left(\frac{x}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right).$$

Show that F is the standard normal distribution.

28.15. (a) Let Y_{nk} be independent random variables having the Poisson distribution with mean $cn^\alpha/|k|^{1+\alpha}$, where $c > 0$ and $0 < \alpha < 2$. Let $Z_n = n^{-1}\sum_{k=-n^2}^{n^2} kY_{nk}$ (omit $k = 0$ in the sum), and show that if c is properly chosen then the characteristic function of Z_n converges to $e^{-|t|^\alpha}$.

(b) Show for $0 < \alpha \le 2$ that $e^{-|t|^\alpha}$ is the characteristic function of a symmetric stable distribution; it is called the *symmetric stable law of exponent* α. The case $\alpha = 2$ is the normal law, and $\alpha = 1$ is the Cauchy law.

SECTION 29. LIMIT THEOREMS IN R^k

If F_n and F are distribution functions on R^k, then F_n *converges weakly* to F, written $F_n \Rightarrow F$, if $\lim_n F_n(x) = F(x)$ for all continuity points x of F. The corresponding distributions μ_n and μ are in this case also said to converge weakly: $\mu_n \Rightarrow \mu$. If X_n and X are k-dimensional random vectors (possibly on different probability spaces), X_n *converges in distribution* to X, written $X_n \Rightarrow X$, if the corresponding distribution functions converge weakly. The definitions are thus exactly as for the line.

The Basic Theorems

The closure A^- of a set in R^k is the set of limits of sequences in A; the interior is $A° = R^k - (R^k - A)^-$; and the boundary is $\partial A = A^- - A°$. A Borel set A is a *μ-continuity set* if $\mu(\partial A) = 0$. The first theorem is the k-dimensional version of Theorem 25.8.

Theorem 29.1. *For probability measures μ_n and μ on $(R^k, \mathcal{R}^k)$, each of the following conditions is equivalent to the weak convergence of μ_n to μ:*

(i) $\lim_n \int f d\mu_n = \int f d\mu$ *for bounded continuous f*;

(ii) $\limsup_n \mu_n(C) \le \mu(C)$ *for closed C*;

(iii) $\liminf_n \mu_n(G) \ge \mu(G)$ *for open G*;

(iv) $\lim_n \mu_n(A) = \mu(A)$ *for μ-continuity sets A*.

PROOF. It will first be shown that (i) through (iv) are all equivalent.

(i) *implies* (ii): Consider the distance $\text{dist}(x, C) = \inf\{|x - y|: y \in C\}$ from x to C. It is continuous in x. Let

$$\varphi_j(t) = \begin{cases} 1 & \text{if } t \le 0, \\ 1 - jt & \text{if } 0 \le t \le j^{-1}, \\ 0 & \text{if } j^{-1} \le t. \end{cases}$$

Then $f_j(x) = \varphi_j(\text{dist}(x,C))$ is continuous and bounded by 1, and $f_j(x) \downarrow I_C(x)$ as $j \uparrow \infty$ because C is closed. If (i) holds, then $\limsup_n \mu_n(C) \le \lim_n \int f_j \, d\mu_n = \int f_j \, d\mu$. As $j \uparrow \infty$, $\int f_j \, d\mu \downarrow \int I_C \, d\mu = \mu(C)$.

(ii) *is equivalent to* (iii). Take $C = R^k - G$.

(ii) *and* (iii) *imply* (iv): From (ii) and (iii) follows

$$\mu(A^\circ) \le \liminf_n \mu_n(A^\circ) \le \liminf_n \mu_n(A)$$

$$\le \limsup_n \mu_n(A) \le \limsup_n \mu_n(A^-) \le \mu(A^-).$$

Clearly (iv) follows from this.

(iv) *implies* (i): Suppose that f is continuous and $|f(x)|$ is bounded by K. Given ϵ, choose reals $\alpha_0 < \alpha_1 < \cdots < \alpha_l$ so that $\alpha_0 < -K < K < \alpha_l$, $\alpha_i - \alpha_{i-1} < \epsilon$, and $\mu[x: f(x) = \alpha_i] = 0$. The last condition can be achieved because the sets $[x: f(x) = \alpha]$ are disjoint for different α. Put $A_i = [x: \alpha_{i-1} < f(x) \le \alpha_i]$. Since f is continuous, $A_i^- \subset [x: \alpha_{i-1} \le f(x) \le \alpha_i]$ and $A_i^\circ \supset [x: \alpha_{i-1} < f(x) < \alpha_i]$. Therefore, $\partial A_i \subset [x: f(x) = \alpha_{i-1}] \cup [x: f(x) = \alpha_i]$, and therefore $\mu(\partial A_i) = 0$. Now $|\int f \, d\mu_n - \sum_{i=1}^l \alpha_i \mu_n(A_i)| \le \epsilon$ and similarly for μ, and $\sum_{i=1}^l \alpha_i \mu_n(A_i) \to \sum_{i=1}^l \alpha_i \mu(A_i)$ because of (iv). Since ϵ was arbitrary, (i) follows.

It remains to prove these four conditions equivalent to weak convergence.

(iv) *implies* $\mu_n \Rightarrow \mu$: Consider the corresponding distribution functions. If $S_x = [y: y_i \le x_i, \ i = 1, \ldots, k]$, then F is continuous at x if and only if $\mu(\partial S_x) = 0$; see the argument following (20.18). Therefore, if F is continuous at x, $F_n(x) = \mu_n(S_x) \to \mu(S_x) = F(x)$, and $F_n \Rightarrow F$.

$\mu_n \Rightarrow \mu$ *implies* (iii): Since only countably many parallel hyperplanes can have positive μ-measure, there is a dense set D of reals such that $\mu[x: x_i = d] = 0$ for $d \in D$ and $i = 1, \ldots, k$. Let $\mathscr{A}$ be the class of rectangles $A = [x: a_i < x_i \le b_i, \ i = 1, \ldots, k]$ for which the a_i and the b_i all lie in D. All 2^k vertices of such a rectangle are continuity points of F, and so $F_n \Rightarrow F$ implies (see (12.12)) that $\mu_n(A) = \Delta_A F_n \to \Delta_A F = \mu(A)$. It follows by the inclusion–exclusion formula that $\mu_n(B) \to \mu(B)$ for finite unions B of elements of $\mathscr{A}$. Since D is dense on the line, an open set G in R^k is a countable union of sets A_m in $\mathscr{A}$. But $\mu(\bigcup_{m \le M} A_m) = \lim_n \mu_n(\bigcup_{m \le M} A_m) \le \liminf_n \mu_n(G)$. Letting $M \to \infty$ gives (iii). ∎

Theorem 29.2. *Suppose that* $h: R^k \to R^j$ *is measurable and that the set* D_h *of its discontinuities is measurable.*[†] *If* $\mu_n \Rightarrow \mu$ *in* R^k *and* $\mu(D_h) = 0$, *then* $\mu_n h^{-1} \Rightarrow \mu h^{-1}$ *in* R^j.

[†]The argument in the footnote on p. 334 shows that in fact $D_h \in \mathscr{R}^k$ always holds.

PROOF. Let C be a closed set in R^j. The closure $(h^{-1}C)^-$ in R^k satisfies $(h^{-1}C)^- \subset D_h \cup h^{-1}C$. If $\mu_n \Rightarrow \mu$, then part (ii) of Theorem 29.1 gives

$$\lim_n \sup \mu_n h^{-1}(C) \leq \lim_n \sup \mu_n\big((h^{-1}C)^-\big) \leq \mu\big((h^{-1}C)^-\big)$$

$$\leq \mu(D_h) + \mu\big(h^{-1}C\big).$$

Using (ii) again gives $\mu_n h^{-1} \Rightarrow \mu h^{-1}$ if $\mu(D_h) = 0$. ∎

Theorem 29.2 is the k-dimensional version of the *mapping theorem*—Theorem 25.7. The two proofs just given provide in the case $k = 1$ a second approach to the theory of Section 25, which there was based on Skorohod's theorem (Theorem 25.6). Skorohod's theorem does extend to R^k, but the proof is harder.[†]

Theorems 29.1 and 29.2 can of course be stated in terms of random vectors. For example, $X_n \Rightarrow X$ if and only if $P[X \in G] \leq \liminf_n P[X_n \in G]$ for all open sets G.

A sequence $\{\mu_n\}$ of probability measures on $(R^k, \mathscr{R}^k)$ is *tight* if for every ϵ there is a bounded rectangle A such that $\mu_n(A) > 1 - \epsilon$ for all n.

Theorem 29.3. *If $\{\mu_n\}$ is a tight sequence of probability measures, there is a subsequence $\{\mu_{n_i}\}$ and a probability measure μ such that $\mu_{n_i} \Rightarrow \mu$ as $i \to \infty$.*

PROOF. Take $S_x = [y: y_j \leq x_j, j \leq k]$ and $F_n(x) = \mu_n(S_x)$. The proof of Helly's theorem (Theorem 25.9) carries over: For points x and y in R^k, interpret $x \leq y$ as meaning $x_u \leq y_u$, $u = 1, \ldots, k$, and $x < y$ as meaning $x_u < y_u$, $u = 1, \ldots, k$. Consider rational points r—points whose coordinates are all rational—and by the diagonal method [A14] choose a sequence $\{n_i\}$ along which $\lim_i F_{n_i}(r) = G(r)$ exists for each such r. As before, define $F(x) = \inf[G(r): x < r]$. Although F is clearly nondecreasing in each variable, a further argument is required to prove $\Delta_A F \geq 0$ (see (12.12)).

Given ϵ and a rectangle $A = (a_1, b_1] \times \cdots \times (a_k, b_k]$, choose a δ such that if $z = (\delta, \ldots, \delta)$, then for each of the 2^k vertices x of A, $x < r < x + z$ implies $|F(x) - G(r)| < \epsilon/2^k$. Now choose rational points r and s such that $a < r < a + z$ and $b < s < b + z$. If $B = (r_1, s_1] \times \cdots \times (r_k, s_k]$, then $|\Delta_A F - \Delta_B G| < \epsilon$. Since $\Delta_B G = \lim_i \Delta_B F_{n_i} \geq 0$ and ϵ is arbitrary, it follows that $\Delta_A F \geq 0$.

With the present interpretation of the symbols, the proof of Theorem 25.9 shows that F is continuous from above and $\lim_i F_{n_i}(x) = F(x)$ for continuity points x of F.

[†]The approach of this section carries over to general metric spaces; for this theory and its applications, see BILLINGSLEY₁ and BILLINGSLEY₂. Since Skorohod's theorem is no easier in R^k than in the general metric space, it is not treated here.

By Theorem 12.5, there is a measure μ on $(R^k, \mathscr{R}^k)$ such that $\mu(A) = \Delta_A F$ for rectangles A. By tightness, there is for given ϵ a t such that $\mu_n[y: -t < y_j \leq t, \, j \leq k] > 1 - \epsilon$ for all n. Suppose that all coordinates of x exceed t: If $r > x$, then $F_n(r) > 1 - \epsilon$ and hence (r rational) $G(r) \geq 1 - \epsilon$, so that $F(x) \geq 1 - \epsilon$. Suppose, on the other hand, that some coordinate of x is less than $-t$: Choose a rational r such that $x < r$ and some coordinate of r is less than $-t$; then $F_n(r) < \epsilon$, hence $G(r) \leq \epsilon$, and so $F(x) \leq \epsilon$. Therefore, for every ϵ there is a t such that

$$(29.1) \qquad F(x) \begin{cases} \geq 1 - \epsilon & \text{if } x_j > t \text{ for all } j, \\ \leq \epsilon & \text{if } x_j < -t \text{ for some } j. \end{cases}$$

If $B_s = [y: -s < y_j \leq x_j, \, j \leq k]$, then $\mu(S_x) = \lim_s \mu(B_s) = \lim_s \Delta_{B_s} F$. Of the 2^k terms in the sum $\Delta_{B_s} F$, all but $F(x)$ go to 0 ($s \to \infty$) because of the second part of (29.1). Thus $\mu(S_x) = F(x)$.[†] Because of the other part of (29.1), μ is a probability measure. Therefore, $F_n \Rightarrow F$ and $\mu_n \Rightarrow \mu$. ∎

Obviously Theorem 29.3 implies that tightness is a sufficient condition that each subsequence of $\{\mu_n\}$ contain a further subsequence converging weakly to some probability measure. (An easy modification of the proof of Theorem 25.10 shows that tightness is necessary for this as well.) And clearly the corollary to Theorem 25.10 now goes through:

Corollary. *If $\{\mu_n\}$ is a tight sequence of probability measures, and if each subsequence that converges weakly at all converges weakly to the probability measure μ, then $\mu_n \Rightarrow \mu$.*

Characteristic Functions

Consider a random vector $X = (X_1, \ldots, X_k)$ and its distribution μ in R^k. Let $t \cdot x = \sum_{u=1}^k t_u x_u$ denote inner product. The characteristic function of X and of μ is defined over R^k by

$$(29.2) \qquad \varphi(t) = \int_{R^k} e^{it \cdot x} \mu(dx) = E[e^{it \cdot X}].$$

To a great extent its properties parallel those of the one-dimensional characteristic function and can be deduced by parallel arguments.

[†] This requires proof because there exist (Problem 12.10) functions F' other than F for which $\mu(A) = \Delta_A F'$ holds for all rectangles A.

The inversion formula (26.16) takes this form: For a bounded rectangle $A = [x: a_u < x_u \le b_u, u \le k]$ such that $\mu(\partial A) = 0$,

$$(29.3) \qquad \mu(A) = \lim_{T \to \infty} \frac{1}{(2\pi)^k} \int_{B_T} \prod_{u=1}^{k} \frac{e^{-it_u a_u} - e^{-it_u b_u}}{it_u} \varphi(t) \, dt,$$

where $B_T = [t \in R^k: |t_u| \le T, u \le k]$ and dt is short for $dt_1 \cdots dt_k$. To prove it, replace $\varphi(t)$ by the middle term in (29.2) and reverse the integrals as in (26.17): The integral in (29.3) is

$$I_T = \frac{1}{(2\pi)^k} \int_{R^k} \left[\int_{B_T} \prod_{u=1}^{k} \frac{e^{-it_u a_u} - e^{-it_u b_u}}{it_u} e^{it_u x_u} \, dt \right] \mu(dx).$$

The inner integral may be evaluated by Fubini's theorem in R^k, which gives

$$I_T = \int_{R^k} \prod_{u=1}^{k} \left[\frac{\mathrm{sgn}(x_u - a_u)}{\pi} S(T \cdot |x_u - a_u|) \right.$$
$$\left. - \frac{\mathrm{sgn}(x_u - b_u)}{\pi} S(T \cdot |x_u - b_u|) \right] \mu(dx).$$

Since the integrand converges to $\prod_{u=1}^{k} \psi_{a_u, b_u}(x_u)$ (see (26.18)), (29.3) follows as in the case $k = 1$.

The proof that weak convergence implies (iii) in Theorem 29.1 shows that for probability measures μ and ν on R^k there exists a dense set D of reals such that $\mu(\partial A) = \nu(\partial A) = 0$ for all rectangles A whose vertices have coordinates in D. If $\mu(A) = \nu(A)$ for such rectangles, then μ and ν are identical by Theorem 3.3.

Thus *the characteristic function φ uniquely determines the probability measure μ* Further properties of the characteristic function can be derived from the one-dimensional case by means of the following device of Cramér and Wold. For $t \in R^k$, define $h_t: R^k \to R^1$ by $h_t(x) = t \cdot x$. For real α, $[x: t \cdot x \le \alpha]$ is a half space, and its μ-measure is

$$(29.4) \qquad \mu[x: t \cdot x \le \alpha] = \mu h_t^{-1}(-\infty, \alpha].$$

By change of variable, the characteristic function of μh_t^{-1} is

$$(29.5) \qquad \int_{R^1} e^{isy} \mu h_t^{-1}(dy) = \int_{R^k} e^{is(t \cdot x)} \mu(dx)$$
$$= \varphi(st_1, \ldots, st_k), \qquad s \in R^1.$$

To know the μ-measure of every half space is (by (29.4)) to know each μh_t^{-1} and hence is (by (29.5) for $s = 1$) to know $\varphi(t)$ for every t; and to know the

characteristic function φ of μ is to know μ. Thus μ *is uniquely determined by the values it gives to the half spaces.* This result, very simple in its statement, seems to require Fourier methods—no elementary proof is known.

If $\mu_n \Rightarrow \mu$ for probability measures on R^k, then $\varphi_n(t) \to \varphi(t)$ for the corresponding characteristic functions by Theorem 29.1. But suppose that the characteristic functions converge pointwise. It follows by (29.5) that for each t the characteristic function of $\mu_n h_t^{-1}$ converges pointwise on the line to the characteristic function of μh_t^{-1}; by the continuity theorem for characteristic functions on the line then, $\mu_n h_t^{-1} \Rightarrow \mu h_t^{-1}$. Take the uth component of t to be 1 and the others 0; then the $\mu_n h_t^{-1}$ are the marginals for the uth coordinate. Since $\{\mu_n h_t^{-1}\}$ is weakly convergent, there is a bounded interval $(a_u, b_u]$ such that $\mu_n[x \in R^k: a_u < x_u \le b_u] = \mu_n h_t^{-1}(a_u, b_u] > 1 - \epsilon/k$ for all n. But then $\mu_n(A) > 1 - \epsilon$ for the bounded rectangle $A = [x: a_u < x_u \le b_u, u = 1, \ldots, k]$. The sequence $\{\mu_n\}$ is therefore tight. If a subsequence $\{\mu_{n_i}\}$ converges weakly to v, then $\varphi_{n_i}(t)$ converges to the characteristic function of v, which is therefore $\varphi(t)$. By uniqueness, $v = \mu$, so that $\mu_{n_i} \Rightarrow \mu$. By the corollary to Theorem 29.3, $\mu_n \Rightarrow \mu$. This proves the continuity theorem for k-dimensional characteristic functions: $\mu_n \Rightarrow \mu$ *if and only if* $\varphi_n(t) \to \varphi(t)$ *for all t.*

The Cramér–Wold idea leads also to the following result, by means of which certain limit theorems can be reduced in a routine way to the one-dimensional case.

Theorem 29.4. *For random vectors* $X_n = (X_{n1}, \ldots, X_{nk})$ *and* $Y = (Y_1, \ldots, Y_k)$, *a necessary and sufficient condition for* $X_n \Rightarrow Y$ *is that* $\sum_{u=1}^k t_u X_{nu} \Rightarrow \sum_{u=1}^k t_u Y_u$ *for each* $(t_1, \ldots, t_k)$ *in* R^k.

PROOF. The necessity follows from a consideration of the continuous mapping h_t above—use Theorem 29.2. As for sufficiency, the condition implies by the continuity theorem for one-dimensional characteristic functions that for each $(t_1, \ldots, t_k)$

$$E\left[e^{is\sum_{u=1}^k t_u X_{nu}}\right] \to E\left[e^{is\sum_{u=1}^k t_u Y_u}\right]$$

for all real s. Taking $s = 1$ shows that the characteristic function of X_n converges pointwise to that of Y. ∎

Normal Distributions in R^k

By Theorem 20.4 there is (on some probability space) a random vector $X = (X_1, \ldots, X_k)$ with independent components each having the standard normal distribution. Since each X_u has density $e^{-x^2/2}/\sqrt{2\pi}$, X has density

(see (20.25))

$$(29.6) \qquad f(x) = \frac{1}{(2\pi)^{k/2}} e^{-|x|^2/2},$$

where $|x|^2 = \sum_{u=1}^{k} x_u^2$ denotes Euclidean norm. This distribution plays the role of the standard normal distribution in R^k. Its characteristic function is

$$(29.7) \qquad E\left[\prod_{u=1}^{k} e^{it_u X_u}\right] = \prod_{u=1}^{k} e^{-t_u^2/2} = e^{-|t|^2/2}.$$

Let $A = [a_{ui}]$ be a $k \times k$ matrix, and put $Y = AX$, where X is viewed as a column vector. Since $E[X_\alpha X_\beta] = \delta_{\alpha\beta}$, the matrix $\Sigma = [\sigma_{ui}]$ of the covariances of Y has entries $\sigma_{ui} = E[Y_u Y_i] = \sum_{\alpha=1}^{k} a_{u\alpha} a_{i\alpha}$. Thus $\Sigma = AA'$, where the prime denotes transpose. The matrix Σ is symmetric and nonnegative definite: $\sum_{ui} \sigma_{ui} x_u x_i = |A'x|^2 \geq 0$. View t also as a column vector with transpose t', and note that $t \cdot x = t'x$. The characteristic function of AX is thus

$$(29.8) \qquad E[e^{it'(AX)}] = E[e^{i(A't)'X}] = e^{-|A't|^2/2} = e^{-t'\Sigma t/2}.$$

Define a centered *normal distribution* as any probability measure whose characteristic function has this form for some symmetric nonnegative definite Σ.

If Σ is symmetric and nonnegative definite, then for an appropriate orthogonal matrix U, $U'\Sigma U = D$ is a diagonal matrix whose diagonal elements are the eigenvalues of Σ and hence are nonnegative. If D_0 is the diagonal matrix whose elements are the square roots of those of D, and if $A = UD_0$, then $\Sigma = AA'$. Thus for every nonnegative definite Σ there exists a centered normal distribution (namely the distribution of AX) with covariance matrix Σ and characteristic function $\exp(-\frac{1}{2}t'\Sigma t)$.

If Σ is nonsingular, so is the A just constructed. Since X has density (29.6), $Y = AX$ has, by the Jacobian transformation formula (20.20), density $f(A^{-1}x)|\det A^{-1}|$. From $\Sigma = AA'$ follows $|\det A^{-1}| = (\det \Sigma)^{-1/2}$. Moreover, $\Sigma^{-1} = (A')^{-1}A^{-1}$, so that $|A^{-1}x|^2 = x'\Sigma^{-1}x$. Thus the normal distribution has density $(2\pi)^{k/2}(\det \Sigma)^{-1/2}\exp(-\frac{1}{2}x'\Sigma^{-1}x)$ if Σ is nonsingular. If Σ is singular, the A constructed above must be singular as well, so that AX is confined to some hyperplane of dimension $k - 1$ and the distribution can have no density.

By (29.8) and the uniqueness theorem for characteristic functions in R^k, a *centered normal distribution is completely determined by its covariance matrix.* Suppose the off-diagonal elements of Σ are 0, and let A be the diagonal matrix with the $\sigma_{ii}^{1/2}$ along the diagonal. Then $\Sigma = AA'$, and if X has the standard normal distribution, the components X_i are independent and hence so are the components $\sigma_{ii}^{1/2}X_i$ of AX. Therefore, *the components of a*

normally distributed random vector are independent if and only if they are uncorrelated.

If M is a $j \times k$ matrix and Y has in R^k the centered normal distribution with covariance matrix Σ, then MY has in R^j the characteristic function $\exp(-\frac{1}{2}(M't)'\Sigma(M't)) = \exp(-\frac{1}{2}t'(M\Sigma M')t)$ $(t \in R^j)$. Hence MY has the centered normal distribution in R^j with covariance matrix $M\Sigma M'$. Thus *a linear transformation of a normal distribution is itself normal.*

These normal distributions are special in that all the first moments vanish. The general normal distribution is a translation of one of these centered distributions. It is completely determined by its means and covariances.

The Central Limit Theorem

Let $X_n = (X_{n1}, \ldots, X_{nk})$ be independent random vectors all having the same distribution. Suppose that $E[X_{nu}^2] < \infty$; let the vector of means be $c = (c_1, \ldots, c_k)$, where $c_u = E[X_{nu}]$, and let the covariance matrix be $\Sigma = [\sigma_{ul}]$, where $\sigma_{ul} = E[(X_{nu} - c_u)(X_{nl} - c_l)]$. Put $S_n = X_1 + \cdots + X_n$.

Theorem 29.5. *Under these assumptions, the distribution of the random vector $(S_n - nc)/\sqrt{n}$ converges weakly to the centered normal distribution with covariance matrix Σ.*

PROOF. Let $Y = (Y_1, \ldots, Y_k)$ be a normally distributed random vector with 0 means and covariance matrix Σ. For given $t = (t_1, \ldots, t_k)$, let $Z_n = \sum_{u=1}^{k} t_u (X_{nu} - c_u)$ and $Z = \sum_{u=1}^{k} t_u Y_u$. By Theorem 29.4, it suffices to prove that $n^{-1/2} \sum_{j=1}^{n} Z_j \Rightarrow Z$ (for arbitrary t). But this is an instant consequence of the Lindeberg–Lévy theorem (Theorem 27.1). ∎

PROBLEMS

29.1. A real function f on R^k is everywhere upper semicontinuous (see Problem 13.8) if for each x and ϵ there is a δ such that $|x - y| < \delta$ implies that $f(y) < f(x) + \epsilon$; f is lower semicontinuous if $-f$ is upper semicontinuous.

(a) Use condition (iii) of Theorem 29.1, Fatou's lemma, and (21.9) to show that, if $\mu_n \Rightarrow \mu$ and f is bounded and lower semicontinuous, then

$$(29.9) \qquad \liminf_n \int f \, d\mu_n \geq \int f \, d\mu.$$

(b) Show that, if (29.9) holds for all bounded, lower semicontinuous functions f, then $\mu_n \Rightarrow \mu$.

(c) Prove the analogous results for upper semicontinuous functions.

29.2. (a) Show for probability measures on the line that $\mu_n \times \nu_n \Rightarrow \mu \times \nu$ if and only if $\mu_n \Rightarrow \mu$ and $\nu_n \Rightarrow \nu$.

(b) Suppose that X_n and Y_n are independent and that X and Y are independent. Show that, if $X_n \Rightarrow X$ and $Y_n \Rightarrow Y$, then $(X_n, Y_n) \Rightarrow (X, Y)$ and hence that $X_n + Y_n \Rightarrow X + Y$.

(c) Show that part (b) fails without independence.

(d) If $F_n \Rightarrow F$ and $G_n \Rightarrow G$, then $F_n * G_n \Rightarrow F * G$. Prove this by part (b) and also by characteristic functions.

29.3. (a) Show that $\{\mu_n\}$ is tight if and only if for each ϵ there is a compact set K such that $\mu_n(K) > 1 - \epsilon$ for all n.

(b) Show that $\{\mu_n\}$ is tight if and only if each of the k sequences of marginal distributions is tight on the line.

29.4. Assume of (X_n, Y_n) that $X_n \Rightarrow X$ and $Y_n \Rightarrow c$. Show that $(X_n, Y_n) \Rightarrow (X, c)$. This is an example of Problem 29.2(b) where X_n and Y_n need not be assumed independent.

29.5. Prove analogues for R^k of the corollaries to Theorem 26.3.

29.6. Suppose that $f(X)$ and $g(Y)$ are uncorrelated for all bounded continuous f and g. Show that X and Y are independent. *Hint*: Use characteristic functions.

29.7. 20.16↑ Suppose that the random vector X has a centered k-dimensional normal distribution whose covariance matrix has 1 as an eigenvalue of multiplicity r and 0 as an eigenvalue of multiplicity $k - r$. Show that $|X|^2$ has the chi-squared distribution with r degrees of freedom.

29.8. ↑ *Multinomial sampling.* Let $p_1, \ldots, p_k$ be positive and add to 1, and let $Z_1, Z_2, \ldots$ be independent k-dimensional random vectors such that Z_n has with probability p_i a 1 in the ith component and 0's elsewhere. Then $f_n = (f_{n1}, \ldots, f_{nk}) = \sum_{m=1}^{n} Z_m$ is the frequency count for a sample of size n from a multinomial population with cell probabilities p_i. Put $X_{ni} = (f_{ni} - np_i)/\sqrt{np_i}$ and $X_n = (X_{n1}, \ldots, X_{nk})$.

(a) Show that X_n has mean values 0 and covariances $\sigma_{ij} = (\delta_{ij}p_j - p_ip_j)/\sqrt{p_ip_j}$.

(b) Show that the chi squared statistic $\sum_{i=1}^{k}(f_{ni} - np_i)^2/np_i$ has asymptotically the chi-squared distribution with $k - 1$ degrees of freedom.

29.9. 20.26↑ *A theorem of Poincaré.* (a) Suppose that $X_n = (X_{n1}, \ldots, X_{nn})$ is uniformly distributed over the surface of a sphere of radius $\sqrt{n}$ in R^n. Fix t, and show that $X_{n1}, \ldots, X_{nt}$ are in the limit independent, each with the standard normal distribution. *Hint*: If the components of $Y_n = (Y_{n1}, \ldots, Y_{nn})$ are independent, each with the standard normal distribution, then X_n has the same distribution as $\sqrt{n}\, Y_n/|Y_n|$.

(b) Suppose that the distribution of $X_n = (X_{n1}, \ldots, X_{nn})$ is spherically symmetric in the sense that $X_n/|X_n|$ is uniformly distributed over the unit sphere. Assume that $|X_n|^2/n \Rightarrow 1$, and show that $X_{n1}, \ldots, X_{nt}$ are asymptotically independent and normal.

29.10. Let $X_n = (X_{n1}, \ldots, X_{nk})$, $n = 1, 2, \ldots$, be random vectors satisfying the mixing condition (27.19) with $\alpha_n = O(n^{-5})$. Suppose that the sequence is stationary (the distribution of $(X_n, \ldots, X_{n+j})$ is the same for all n), that $E[X_{nu}] = 0$, and that the X_{nu} are uniformly bounded. Show that if $S_n = X_1 + \cdots + X_n$, then $S_n/\sqrt{n}$ has in the limit the centered normal distribution with covariances

$$E[X_{1u}X_{1t}] + \sum_{j=1}^{\infty} E[X_{1u}X_{1+j,t}] + \sum_{j=1}^{\infty} E[X_{1+j,u}X_{1t}].$$

Hint: Use the Cramér–Wold device.

29.11. ↑ As in Example 27.6, let $\{Y_n\}$ be a Markov chain with finite state space $S = \{1, \ldots, s\}$, say. Suppose the transition probabilities p_{ut} are all positive and the initial probabilities p_u are the stationary ones. Let f_{nu} be the number of i for which $1 \leq i \leq n$ and $Y_i = u$. Show that the normalized frequency count

$$n^{-1/2}(f_{n1} - np_1, \ldots, f_{nk} - np_k)$$

has in the limit the centered normal distribution with covariances

$$\delta_{ut} - p_u p_t + \sum_{j=1}^{\infty} \left(p_{ut}^{(j)} - p_u p_t \right) + \sum_{j=1}^{\infty} \left(p_{tu}^{(j)} - p_t p_u \right).$$

29.12. Assume that

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}$$

is positive definite, invert it explicitly, and show that the corresponding two-dimensional normal density is

$$(29.10) \qquad f(x_1, x_2) = \frac{1}{2\pi D^{1/2}} \exp\left[-\frac{1}{2D} \left(\sigma_{22} x_1^2 - 2\sigma_{12} x_1 x_2 + \sigma_{11} x_2^2 \right) \right],$$

where $D = \sigma_{11}\sigma_{22} - \sigma_{12}^2$.

29.13. Suppose that Z has the standard normal distribution in R^1. Let μ be the mixture with equal weights of the distributions of (Z, Z) and $(Z, -Z)$, and let (X, Y) have distribution μ. Prove:

(a) Although each of X and Y is normal, they are not jointly normal.

(b) Although X and Y are uncorrelated, they are not independent.

SECTION 30. THE METHOD OF MOMENTS*

The Moment Problem

For some distributions the characteristic function is intractable but moments can nonetheless be calculated. In these cases it is sometimes possible to prove weak convergence of the distributions by establishing that the moments converge. This approach requires conditions under which a distribution is uniquely determined by its moments, and this is for the same reason that the continuity theorem for characteristic functions requires for its proof the uniqueness theorem.

Theorem 30.1. *Let μ be a probability measure on the line having finite moments $\alpha_k = \int_{-\infty}^{\infty} x^k \mu(dx)$ of all orders. If the power series $\sum_k \alpha_k r^k / k!$ has a positive radius of convergence, then μ is the only probability measure with the moments $\alpha_1, \alpha_2, \ldots$.*

Proof. Let $\beta_k = \int_{-\infty}^{\infty} |x|^k \mu(dx)$ be the absolute moments. The first step is to show that

$$(30.1) \qquad \frac{\beta_k r^k}{k!} \to 0, \qquad k \to \infty,$$

for some positive r. By hypothesis there exists an s, $0 < s < 1$, such that $\alpha_k s^k / k! \to 0$. Choose $0 < r < s$; then $2k r^{2k-1} < s^{2k}$ for large k. Since $|x|^{2k-1} \le 1 + |x|^{2k}$,

$$\frac{\beta_{2k-1} r^{2k-1}}{(2k-1)!} \le \frac{r^{2k-1}}{(2k-1)!} + \frac{\beta_{2k} s^{2k}}{(2k)!}$$

for large k. Hence (30.1) holds as k goes to infinity through odd values; since $\beta_k = \alpha_k$ for k even, (30.1) follows.

By (26.4),

$$\left| e^{itx} \left(e^{ihx} - \sum_{k=0}^{n} \frac{(ihx)^k}{k!} \right) \right| \le \frac{|hx|^{n+1}}{(n+1)!},$$

and therefore the characteristic function φ of μ satisfies

$$\left| \varphi(t+h) - \sum_{k=0}^{n} \frac{h^k}{k!} \int_{-\infty}^{\infty} (ix)^k e^{itx} \mu(dx) \right| \le \frac{|h|^{n+1} \beta_{n+1}}{(n+1)!}.$$

---

*This section may be omitted.

By (26.10), the integral here is $\varphi^{(k)}(t)$. By (30.1),

$$(30.2) \qquad \varphi(t+h) = \sum_{k=0}^{\infty} \frac{\varphi^{(k)}(t)}{k!} h^k, \qquad |h| \le r.$$

If ν is another probability measure with moments α_k and characteristic function $\psi(t)$, the same argument gives

$$(30.3) \qquad \psi(t+h) = \sum_{k=0}^{\infty} \frac{\psi^{(k)}(t)}{k!} h^k, \qquad |h| \le r.$$

Take $t = 0$; since $\varphi^{(k)}(0) = i^k \alpha_k = \psi^{(k)}(0)$ (see (26.9)), φ and ψ agree in $(-r, r)$ and hence have identical derivatives there. Taking $t = r - \epsilon$ and $t = -r + \epsilon$ in (30.2) and (30.3) shows that φ and ψ also agree in $(-2r + \epsilon, 2r - \epsilon)$ and hence in $(-2r, 2r)$. But then they must by the same argument agree in $(-3r, 3r)$ as well, and so on.[†] Thus φ and ψ coincide, and by the uniqueness theorem for characteristic functions, so do μ and ν. ∎

A probability measure satisfying the conclusion of the theorem is said to be *determined by its moments*.

Example 30.1. For the standard normal distribution, $|\alpha_k| \le k!$, and so the theorem implies that it is determined by its moments. ∎

But not all measures are determined by their moments:

Example 30.2. If N has the standard normal density, then e^N has the log-normal density

$$f(x) = \begin{cases} \dfrac{1}{\sqrt{2\pi}} \dfrac{1}{x} e^{-(\log x)^2/2} & \text{if } x > 0, \\ 0 & \text{if } x \le 0. \end{cases}$$

Put $g(x) = f(x)(1 + \sin(2\pi \log x))$. If

$$\int_0^{\infty} x^k f(x) \sin(2\pi \log x) \, dx = 0, \qquad k = 0, 1, 2, \ldots,$$

then g, which is nonnegative, will be a probability density and will have the same moments as f. But a change of variable $\log x = s + k$ reduces the

[†] This process is a version of analytic continuation.

integral above to

$$\frac{1}{\sqrt{2\pi}} e^{k^2/2} \int_{-\infty}^{\infty} e^{-s^2/2} \sin 2\pi s \, ds,$$

which vanishes because the integrand is odd.   ■

Theorem 30.2. *Suppose that the distribution of X is determined by its moments, that the X_n have moments of all orders, and that $\lim_n E[X_n^r] = E[X^r]$ for $r = 1, 2, \ldots$. Then $X_n \Rightarrow X$.*

PROOF. Let μ_n and μ be the distributions of X_n and X. Since $E[X_n^2]$ converges, it is bounded, say by K. By Markov's inequality, $P[|X_n| \geq x] \leq K/x^2$, which implies that the sequence $\{\mu_n\}$ is tight.

Suppose that $\mu_{n_k} \Rightarrow \nu$, and let Y be a random variable with distribution ν. If u is an even integer exceeding r, the convergence and hence boundedness of $E[X_n^u]$ implies that $E[X_{n_k}^r] \to E[Y^r]$, by the corollary to Theorem 25.12. By the hypothesis, then, $E[Y^r] = E[X^r]$—that is, ν and μ have the same moments. Since μ is by hypothesis determined by its moments, ν must be the same as μ, and so $\mu_{n_k} \Rightarrow \mu$. The conclusion now follows by the corollary to Theorem 25.10.   ■

Convergence to the log-normal distribution cannot be proved by establishing convergence of moments (take X to have density f and the X_n to have density g in Example 30.2). Because of Example 30.1, however, this approach will work for a normal limit.

Moment Generating Functions

Suppose that μ has a moment generating function $M(s)$ for $s \in [-s_0, s_0]$, $s_0 > 0$. By (21.22), the hypothesis of Theorem 30.1 is satisfied, and so μ is determined by its moments, which are in turn determined by $M(s)$ via (21.23). Thus μ *is determined by* $M(s)$ *if it exists in a neighborhood of* 0.[†] The version of this for one-sided transforms was proved in Section 22—see Theorem 22.2.

Suppose that μ_n and μ have moment generating functions in a common interval $[-s_0, s_0]$, $s_0 > 0$, and suppose that $M_n(s) \to M(s)$ in this interval. Since $\mu_n[(-a, a)^c] \leq e^{-s_0 a}(M_n(-s_0) + M_n(s_0))$, it follows easily that $\{\mu_n\}$ is tight. Since $M(s)$ determines μ, the usual argument now gives $\mu_n \Rightarrow \mu$.

[†]For another proof, see Problem 26.7. The present proof does not require the idea of analyticity.

Central Limit Theorem by Moments

To understand the application of the method of moments, consider once again a sum $S_n = X_{n1} + \cdots + X_{nk_n}$, where $X_{n1}, \ldots, X_{nk_n}$ are independent and

$$(30.4) \qquad E[X_{nk}] = 0, \qquad E[X_{nk}^2] = \sigma_{nk}^2, \qquad s_n^2 = \sum_{k=1}^{k_n} \sigma_{nk}^2.$$

Suppose further that for each n there is an M_n such that $|X_{nk}| \le M_n$, $k = 1, \ldots, k_n$, with probability 1. Finally, suppose that

$$(30.5) \qquad \frac{M_n}{s_n} \to 0.$$

All moments exist, and[†]

$$(30.6) \qquad S_n^r = \sum_{u=1}^{r} \sum{}' \frac{r!}{r_1! \cdots r_u!} \frac{1}{u!} \sum{}'' X_{ni_1}^{r_1} \cdots X_{ni_u}^{r_u},$$

where $\sum'$ extends over the u-tuples $(r_1, \ldots, r_u)$ of positive integers satisfying $r_1 + \cdots + r_u = r$ and $\sum''$ extends over the u-tuples $(i_1, \ldots, i_u)$ of distinct integers in the range $1 \le i_\alpha \le k_n$.

By independence, then,

$$(30.7) \qquad E\left[\left(\frac{S_n}{s_n}\right)^r\right] = \sum_{u=1}^{r} \sum{}' \frac{r!}{r_1! \cdots r_u!} \frac{1}{u!} A_n(r_1, \ldots, r_u),$$

where

$$(30.8) \qquad A_n(r_1, \ldots, r_u) = \sum{}'' \frac{1}{s_n^r} E[X_{ni_1}^{r_1}] \cdots E[X_{ni_u}^{r_u}],$$

and $\sum'$ and $\sum''$ have the same ranges as before. To prove that (30.7) converges to the rth moment of the standard normal distribution, it suffices to show that

$$(30.9) \qquad \lim_n A_n(r_1, \ldots, r_u) = \begin{cases} 1 & \text{if } r_1 = \cdots = r_u = 2, \\ 0 & \text{otherwise} \end{cases}.$$

Indeed, if r is even, all terms in (30.7) will then go to 0 except the one for which $u = r/2$ and $r_\alpha \equiv 2$, which will go to $r!/(r_1! \cdots r_u! u!) = 1 \times 3 \times 5 \times \cdots \times (r-1)$. And if r is odd, the terms will go to 0 without exception.

[†]To deduce this from the multinomial formula, restrict the inner sum to u-tuples satisfying $1 \le i_1 < \cdots < i_u \le k_n$ and compensate by striking out the $1/u!$.

If $r_\alpha = 1$ for some α, then (30.9) holds because by (30.4) each summand in (30.8) vanishes. Suppose that $r_\alpha \geq 2$ for each α and $r_\alpha > 2$ for some α. Then $r > 2u$, and since $|E[X_{ni}^{r_\alpha}]| \leq M_n^{(r_\alpha - 2)}\sigma_{ni}^2$, it follows that $A_n(r_1, \ldots, r_u) \leq (M_n/s_n)^{r-2u}A_n(2, \ldots, 2)$. But this goes to 0 because (30.5) holds and because $A_n(2, \ldots, 2)$ is bounded by 1 (it increases to 1 if the sum in (30.8) is enlarged to include all the u-tuples $(i_1, \ldots, i_u)$).

It remains only to check (30.9) for $r_1 = \cdots = r_u = 2$. As just noted, $A_n(2, \ldots, 2)$ is at most 1, and it differs from 1 by $\sum s_n^{-2u}\sigma_{ni_u}^2$, the sum extending over the $(i_1, \ldots, i_u)$ with at least one repeated index. Since $\sigma_{ni}^2 \leq M_n^2$, the terms for example with $i_u = i_{u-1}$ sum to at most $M_n^2 s_n^{-2u}\sum \sigma_{ni_1}^2 \cdots \sigma_{ni_{u-1}}^2 \leq M_n^2 s_n^{-2}$. Thus $1 - A_n(2, \ldots, 2) \leq u^2 M_n^2 s_n^{-2} \to 0$.

This proves that the moments (30 7) converge to those of the normal distribution and hence that $S_n/s_n \Rightarrow N$

Application to Sampling Theory

Suppose that n numbers

$$x_{n1}, x_{n2}, \ldots, x_{nn},$$

not necessarily distinct, are associated with the elements of a population of size n. Suppose that these numbers are normalized by the requirement

$$(30.10) \qquad \sum_{h=1}^{n} x_{nh} = 0, \qquad \sum_{h=1}^{n} x_{nh}^2 = 1, \qquad M_n = \max_{h \leq n}|x_{nh}|.$$

An ordered sample $X_{n1}, \ldots, X_{nk_n}$ is taken, where the sampling is without replacement. By (30.10), $E[X_{nk}] = 0$ and $E[X_{nk}^2] = 1/n$. Let $s_n^2 = k_n/n$ be the fraction of the population sampled. If the X_{nk} were independent, which they are not, $S_n = X_{n1} + \cdots + X_{nk_n}$ would have variance s_n^2. If k_n is small in comparison with n, the effects of dependence should be small. It will be shown that $S_n/s_n \Rightarrow N$ if

$$(30.11) \qquad s_n^2 = \frac{k_n}{n} \to 0, \qquad \frac{M_n}{s_n} \to 0, \qquad k_n \to \infty.$$

Since $M_n^2 \geq n^{-1}$ by (30.10), the second condition here in fact implies the third.

The moments again have the form (30.7), but this time $E[X_{ni_1}^{r_1} \cdots X_{ni_u}^{r_u}]$ cannot be factored as in (30.8). On the other hand, this expected value is by symmetry the same for each of the $(k_n)_u = k_n(k_n - 1) \cdots (k_n - u + 1)$ choices of the indices i_α in the sum $\sum''$. Thus

$$A_n(r_1, \ldots, r_u) = \frac{(k_n)_u}{s_n^r}E[X_{n1}^{r_1} \cdots X_{nu}^{r_u}].$$

The problem again is to prove (30.9).

The proof goes by induction on u. Now $A_n(r) = k_n s_n^{-r} n^{-1} \sum_{h=1}^n x_{nh}^r$, so that $A_n(1) = 0$ and $A_n(2) = 1$. If $r \geq 3$, then $|x_{nh}^r| \leq M_n^{r-2} x_{nh}^2$, and so $|A_n(r)| \leq (M_n/s_n)^{r-2} \to 0$ by (30.11).

Next suppose as induction hypothesis that (30.9) holds with $u - 1$ in place of u. Since the sampling is without replacement, $E[X_{n1}^{r_1} \cdots X_{nu}^{r_u}] = \sum x_{nh_1}^{r_1} \cdots x_{nh_u}^{r_u}/(n)_u$, where the summation extends over the u-tuples $(h_1,\ldots,h_u)$ of distinct integers in the range $1 \leq h_\alpha \leq n$. In this last sum enlarge the range by requiring of $(h_1, h_2, \ldots, h_u)$ only that $h_2, \ldots, h_u$ be distinct, and then compensate by subtracting away the terms where $h_1 = h_2$, where $h_1 = h_3$, and so on. The result is

$$
E[X_{n1}^{r_1} \cdots X_{nu}^{r_u}] = \frac{n(n)_{u-1}}{(n)_u} E[X_{n1}^{r_1}] E[X_{n2}^{r_2} \cdots X_{nu}^{r_u}]
$$

$$
- \sum_{\alpha=2}^u \frac{(n)_{u-1}}{(n)_u} E[X_{n2}^{r_2} \cdots X_{n\alpha}^{r_1+r_\alpha} \cdots X_{ru}^{r_u}].
$$

This takes the place of the factorization made possible in (30.8) by the assumed independence there. It gives

$$
A_n(r_1, \ldots, r_u) = \frac{n}{n-u+1} \frac{k_n - u + 1}{k_n} A_n(r_1) A_n(r_2, \ldots, r_u)
$$

$$
- \frac{k_n - u + 1}{n - u + 1} \sum_{\alpha=2}^u A_n(r_2, \ldots, r_1 + r_\alpha, \ldots, r_u).
$$

By the induction hypothesis the last sum is bounded, and the factor in front goes to 0 by (30.11). As for the first term on the right, the factor in front goes to 1. If $r_1 \neq 2$, then $A_n(r_1) \to 0$ and $A_n(r_2, \ldots, r_u)$ is bounded, and so $A_n(r_1, \ldots, r_u) \to 0$. The same holds by symmetry if $r_\alpha \neq 2$ for some α other than 1. If $r_1 = \cdots = r_u = 2$, then $A_n(r_1) = 1$, and $A_n(r_2, \ldots, r_u) \to 1$ by the induction hypothesis.

Thus (30.9) holds in all cases, and $S_n/s_n \Rightarrow N$ follows by the method of moments.

Application to Number Theory

Let $g(m)$ be the number of distinct prime factors of the integer m; for example $g(3^4 \times 5^2) = 2$. Since there are infinitely many primes, $g(m)$ is unbounded above; for the same reason, it drops back to 1 for infinitely many m (for the primes and their powers). Since g fluctuates in an irregular way, it is natural to inquire into its average behavior.

On the space Ω of positive integers, let P_n be the probability measure that places mass $1/n$ at each of $1, 2, \ldots, n$, so that among the first n positive

integers the proportion that are contained in a given set A is just $P_n(A)$. The problem is to study $P_n[m: g(m) \le x]$ for large n.

If $\delta_p(m)$ is 1 or 0 according as the prime p divides m or not, then

$$(30.12) \qquad\qquad g(m) = \sum_p \delta_p(m).$$

Probability theory can be used to investigate this sum because under P_n the $\delta_p(m)$ behave somewhat like independent random variables. If $p_1, \ldots, p_u$ are distinct primes, then by the fundamental theorem of arithmetic, $\delta_{p_1}(m) = \cdots = \delta_{p_u}(m) = 1$—that is, each p_i divides m—if and only if the product $p_1 \cdots p_u$ divides m. The probability under P_n of this is just n^{-1} times the number of m in the range $1 \le m \le n$ that are multiples of $p_1 \cdots p_u$, and this number is the integer part of $n/p_1 \cdots p_u$. Thus

$$(30.13) \qquad P_n\left[m: \delta_{p_i}(m) = 1, i = 1, \ldots, u\right] = \frac{1}{n}\left\lfloor \frac{n}{p_1 \cdots p_u} \right\rfloor$$

for distinct p_i.

Now let X_p be independent random variables (on some probability space, one variable for each prime p) satisfying

$$P[X_p = 1] = \frac{1}{p}, \qquad P[X_p = 0] = 1 - \frac{1}{p}.$$

If $p_1, \ldots, p_u$ are distinct, then

$$(30.14) \qquad P\left[X_{p_i} = 1, i = 1, \ldots, u\right] = \frac{1}{p_1 \cdots p_u}.$$

For fixed $p_1, \ldots, p_u$, (30.13) converges to (30.14) as $n \to \infty$. Thus the behavior of the X_p can serve as a guide to that of the $\delta_p(m)$. If $m \le n$, (30.12) is $\sum_{p \le n} \delta_p(m)$, because no prime exceeding m can divide it. The idea[†] is to compare this sum with the corresponding sum $\sum_{p \le n} X_p$.

This will require from number theory the elementary estimate[‡]

$$(30.15) \qquad\qquad \sum_{p \le x} \frac{1}{p} = \log\log x + O(1).$$

The mean and variance of $\sum_{p \le n} X_p$ are $\sum_{p \le n} p^{-1}$ and $\sum_{p \le n} p^{-1}(1 - p^{-1})$; since $\sum_p p^{-2}$ converges, each of these two sums is asymptotically $\log\log n$.

[†]Compare Problems 2.18, 5.19, and 6.16.
[‡]See, for example, Problem 18.17, or HARDY & WRIGHT, Chapter XXII.

Comparing $\Sigma_{p \le n} \delta_p(m)$ with $\Sigma_{p \le n} X_p$ then leads one to conjecture the *Erdös–Kac central limit theorem for the prime divisor function*:

Theorem 30.3. *For all x,*

$$(30.16) \qquad P_n\left[m: \frac{g(m) - \log\log n}{\sqrt{\log\log n}} \le x\right] \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2}\, du.$$

PROOF. The argument uses the method of moments. The first step is to show that (30.16) is unaffected if the range of p in (30.12) is further restricted. Let $\{\alpha_n\}$ be a sequence going to infinity slowly enough that

$$(30.17) \qquad \frac{\log \alpha_n}{\log n} \to 0$$

but fast enough that

$$(30.18) \qquad \sum_{\alpha_n < p \le n} \frac{1}{p} = o(\log\log n)^{1/2}.$$

Because of (30.15), these two requirements are met if, for example, $\log \alpha_n = (\log n)/\log\log n$.

Now define

$$(30.19) \qquad g_n(m) = \sum_{p \le \alpha_n} \delta_p(m).$$

For a function f of positive integers, let

$$E_n[f] = n^{-1} \sum_{m=1}^{n} f(m)$$

denote its expected value computed with respect to P_n. By (30.13) for $u = 1$,

$$E_n\left[\sum_{p > \alpha_n} \delta_p\right] = \sum_{\alpha_n < p \le n} P_n\left[m: \delta_p(m) = 1\right] \le \sum_{\alpha_n < p \le n} \frac{1}{p}.$$

By (30.18) and Markov's inequality,

$$P_n\left[m: |g(m) - g_n(m)| \ge \epsilon(\log\log n)^{1/2}\right] \to 0.$$

Therefore (Theorem 25.4), (30.16) is unaffected if $g_n(m)$ is substituted for $g(m)$.

Now compare (30.19) with the corresponding sum $S_n = \sum_{p \le \alpha_n} X_p$. The mean and variance of S_n are

$$c_n = \sum_{p \le \alpha_n} \frac{1}{p}, \qquad s_n^2 = \sum_{p \le \alpha_n} \frac{1}{p}\left(1 - \frac{1}{p}\right),$$

and each is $\log\log n + o(\log\log n)^{1/2}$ by (30.18). Thus (see Example 25.8), (30.16) with $g(m)$ replaced as above is equivalent to

$$(30.20) \qquad P_n\left[m: \frac{g_n(m) - c_n}{s_n} \le x\right] \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-u^2/2} \, du.$$

It therefore suffices to prove (30.20).

Since the X_p are bounded, the analysis of the moments (30.7) applies here. The only difference is that the summands in S_n are indexed not by the integers k in the range $k \le k_n$ but by the primes p in the range $p \le \alpha_n$; also, X_p must be replaced by $X_p - p^{-1}$ to center it. Thus the rth moment of $(S_n - c_n)/s_n$ converges to that of the normal distribution, and so (30.20) and (30.16) will follow by the method of moments if it is shown that as $n \to \infty$,

$$(30.21) \qquad E\left[\left(\frac{S_n - c_n}{s_n}\right)^r\right] - E_n\left[\left(\frac{g_n - c_n}{s_n}\right)^r\right] \to 0$$

for each r.

Now $E[S_n^r]$ is the sum

$$(30.22) \qquad \sum_{u=1}^r \sideset{}{'}\sum \frac{r!}{r_1! \cdots r_u!} \frac{1}{u!} \sideset{}{''}\sum E\left[X_{p_1}^{r_1} \cdots X_{p_u}^{r_u}\right],$$

where the range of $\sum'$ is as in (30.6) and (30.7), and $\sum''$ extends over the u-tuples $(p_1, \ldots, p_u)$ of distinct primes not exceeding α_n. Since X_p assumes only the values 0 and 1, from the independence of the X_p and the fact that the p_i are distinct, it follows that the summand in (30.22) is

$$(30.23) \qquad E\left[X_{p_1} \cdots X_{p_u}\right] = \frac{1}{p_1 \cdots p_u}.$$

By the definition (30.19), $E_n[g_n^r]$ is just (30.22) with the summand replaced by $E_n[\delta_{p_1}^{r_1} \cdots \delta_{p_u}^{r_u}]$. Since $\delta_p(m)$ assumes only the values 0 and 1, from (30.13) and the fact that the p_i are distinct, it follows that this summand is

$$(30.24) \qquad E_n\left[\delta_{p_1} \cdots \delta_{p_u}\right] = \frac{1}{n}\left\lfloor \frac{n}{p_1 \cdots p_u} \right\rfloor.$$

But (30.23) and (30.24) differ by at most $1/n$, and hence $E[S_n^r]$ and $E_n[g_n^r]$ differ by at most the sum (30.22) with the summand replaced by $1/n$. Therefore,

$$(30.25) \qquad |E[S_n^r] - E_n[g_n^r]| \le \frac{1}{n}\left(\sum_{p \le \alpha_n} 1\right)^r \le \frac{\alpha_n^r}{n}.$$

Now

$$E\left[(S_n - c_n)^r\right] = \sum_{k=0}^{r} \binom{r}{k} E[S_n^k](-c_n)^{r-k},$$

and $E_n[(g_n - c_n)^r]$ has the analogous expansion. Comparing the two expansions term for term and applying (30.25) shows that

$$(30.26) \qquad \left|E\left[(S_n - c_n)^r\right] - E_n\left[(g_n - c_n)^r\right]\right|$$

$$\le \sum_{k=0}^{r} \binom{r}{k}\frac{\alpha_n^k}{n} c_n^{r-k} = \frac{1}{n}(\alpha_n + c_n)^r.$$

Since $c_n \le \alpha_n$, and since $\alpha_n^r/n \to 0$ by (30.17), (30.21) follows as required. ∎

The method of proof requires passing from (30.12) to (30.19). Without this, the α_n on the right in (30.26) would instead be n, and it would not follow that the difference on the left goes to 0; hence the truncation (30.19) for an α_n small enough to satisfy (30.17). On the other hand, α_n must be large enough to satisfy (30.18), in order that the truncation leave (30.16) unaffected.

PROBLEMS

30.1. From the central limit theorem under the assumption (30.5) get the full Lindeberg theorem by a truncation argument.

30.2. For a sample of size k_n with replacement from a population of size n, the probability of no duplicates is $\prod_{j=0}^{k_n-1}(1 - j/n)$. Under the assumption $k_n/\sqrt{n} \to 0$ in addition to (30.10), deduce the asymptotic normality of S_n by a reduction to the independent case.

30.3. By adapting the proof of (21.24), show that the moment generating function of μ in an arbitrary interval determines μ.

30.4. 25.13 30.3↑ Suppose that the moment generating function of μ_n converges to that of μ in some interval. Show that $\mu_n \Rightarrow \mu$.

30.5. Let μ be a probability measure on R^k for which $\int_{R^k} |x_i|^r \mu(dx) < \infty$ for $i = 1, \ldots, k$ and $r = 1, 2, \ldots$. Consider the cross moments

$$\alpha(r_1, \ldots, r_k) = \int_{R^k} x_1^{r_1} \cdots x_k^{r_k} \mu(dx)$$

for nonnegative integers r_i.

(a) Suppose for each i that

(30.27) $$\sum_r \frac{\theta^r}{r!} \int_{R^k} |x_i|^r \mu(dx)$$

has a positive radius of convergence as a power series in θ. Show that μ is determined by its moments in the sense that, if a probability measure ν satisfies $\alpha(r_1, \ldots, r_k) = \int x_1^{r_1} \cdots x_k^{r_k} \nu(dx)$ for all $r_1, \ldots, r_k$, then ν coincides with μ.

(b) Show that a k-dimensional normal distribution is determined by its moments.

30.6. $\uparrow$ Let μ_n and μ be probability measures on R^k. Suppose that for each i, (30.27) has a positive radius of convergence. Suppose that

$$\int_{R^k} x_1^{r_1} \cdots x_k^{r_k} \mu_n(dx) \to \int_{R^k} x_1^{r_1} \cdots x_k^{r_k} \mu(dx)$$

for all nonnegative integers $r_1, \ldots, r_k$. Show that $\mu_n \Rightarrow \mu$.

30.7. 30.5$\uparrow$ Suppose that X and Y are bounded random variables and that X^m and Y^n are uncorrelated for $m, n = 1, 2, \ldots$. Show that X and Y are independent.

30.8. 26.17 30.6$\uparrow$ (a) In the notation (26.32), show for $\lambda \neq 0$ that

(30.28) $$M[(\cos \lambda x)^r] = \binom{r}{r/2} \frac{1}{2^r}$$

for even r and that the mean is 0 for odd r. It follows by the method of moments that $\cos \lambda x$ has a distribution in the sense of (25.18), and in fact of course the relative measure is

(30.29) $$\rho[x : \cos \lambda x \leq u] = 1 - \frac{1}{\pi} \arccos u, \qquad -1 < u < 1.$$

(b) Suppose that $\lambda_1, \lambda_2, \ldots$ are linearly independent over the field of rationals in the sense that, if $n_1 \lambda_1 + \cdots + n_m \lambda_m = 0$ for integers n_ν, then $n_1 = \cdots = n_m = 0$. Show that

(30.30) $$M\left[\prod_{\nu=1}^k (\cos \lambda_\nu x)^{r_\nu}\right] = \prod_{\nu=1}^k M[(\cos \lambda_\nu x)^{r_\nu}]$$

for nonnegative integers $r_1, \ldots, r_k$.

(c) Let $X_1, X_2, \ldots$ be independent and have the distribution function on the right in (30.29). Show that

(30.31) $$\rho\left[x: \sum_{j=1}^{k} \cos \lambda_j x \le u\right] = P[X_1 + \cdots + X_k \le u].$$

(d) Show that

(30.32) $$\lim_{k \to \infty} \rho\left[x: u_1 < \sqrt{\frac{2}{k}} \sum_{j=1}^{k} \cos \lambda_j x \le u_2\right] = \frac{1}{\sqrt{2\pi}} \int_{u_1}^{u_2} e^{-t^2/2} \, dv.$$

For a signal that is the sum of a large number of pure cosine signals with incommensurable frequencies, (30.32) describes the relative amount of time the signal is between u_1 and u_2.

30.9. 6.16↑ From (30.16), deduce once more the Hardy–Ramanujan theorem (see (6.10)).

30.10. ↑ (a) Prove that (if P_n puts probability $1/n$ at $1, \ldots, n$)

(30.33) $$\lim_n P_n\left[m: \left|\frac{\log\log m - \log\log n}{\sqrt{\log\log n}}\right| \ge \epsilon\right] = 0.$$

(b) From (30.16) deduce that (see (2.35) for the notation)

(30.34) $$D\left[m: \frac{g(m) - \log\log m}{\sqrt{\log\log m}} \le x\right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2} \, du.$$

30.11. ↑ Let $G(m)$ be the number of prime factors in m with multiplicity counted. In the notation of Problem 5.19, $G(m) = \sum_p \alpha_p(m)$.
(a) Show for $k \ge 1$ that $P_n[m: \alpha_p(m) - \delta_p(m) \ge k] \le 1/p^{k+1}$; hence $E_n[\alpha_p - \delta_p] \le 2/p^2$.
(b) Show that $E_n[G - g]$ is bounded.
(c) Deduce from (30.16) that

$$P_n\left[m: \frac{G(m) - \log\log n}{\sqrt{\log\log n}} \le x\right] \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2} \, du.$$

(d) Prove for G the analogue of (30.34).

30.12. ↑ Prove the Hardy–Ramanujan theorem in the form

$$D\left[m: \left|\frac{g(m)}{\log\log m} - 1\right| \ge \epsilon\right] = 0.$$

Prove this with G in place of g.

CHAPTER 6

Derivatives and Conditional Probability

SECTION 31. DERIVATIVES ON THE LINE*

This section on Lebesgue's theory of derivatives for real functions of a real variable serves to introduce the general theory of Radon–Nikodym derivatives, which underlies the modern theory of conditional probability. The results here are interesting in themselves and will be referred to later for purposes of illustration and comparison, but they will not be required in subsequent proofs.

The Fundamental Theorem of Calculus

To what extent are the operations of integration and differentiation inverse to one another? A function F is by definition an *indefinite integral* of another function f on $[a, b]$ if

$$(31.1) \qquad F(x) - F(a) = \int_a^x f(t) \, dt$$

for $a \le x \le b$; F is by definition a *primitive* of f if it has derivative f:

$$(31.2) \qquad F'(x) = f(x)$$

for $a \le x \le b$. According to the *fundamental theorem of calculus* (see (17.5)), these concepts coincide in the case of continuous f:

Theorem 31.1. *Suppose that f is continuous on $[a, b]$.*

(i) *An indefinite integral of f is a primitive of f: if (31.1) holds for all x in $[a, b]$, then so does (31.2).*
(ii) *A primitive of f is an indefinite integral of f: if (31.2) holds for all x in $[a, b]$, then so does (31.1).*

---

*This section may be omitted.

400

A basic problem is to investigate the extent to which this theorem holds if f is not assumed continuous. First consider part (i). Suppose f is integrable, so that the right side of (31.1) makes sense. If f is 0 for $x < m$ and 1 for $x \geq m$ ($a < m < b$), then an F satisfying (31.1) has no derivative at m. It is thus too much to ask that (31.2) hold for all x. On the other hand, according to a famous theorem of Lebesgue, if (31.1) holds for all x, then (31.2) holds almost everywhere—that is, except for x in a set of Lebesgue measure 0. In this section *almost everywhere* will refer to Lebesgue measure only. This result, the most one could hope for, will be proved below (Theorem 31.3).

Now consider part (ii) of Theorem 31.1. Suppose that (31.2) holds almost everywhere, as in Lebesgue's theorem, just stated. Does (31.1) follow? The answer is no: If f is identically 0, and if $F(x)$ is 0 for $x < m$ and 1 for $x \geq m$ ($a < m < b$), then (31.2) holds almost everywhere, but (31.1) fails for $x \geq m$. The question was wrongly posed, and the trouble is not far to seek: If f is integrable and (31.1) holds, then

$$(31.3) \qquad F(x+h) - F(x) = \int_a^b I_{(x,x+h)}(t)f(t)\, dt \to 0$$

as $h \downarrow 0$ by the dominated convergence theorem. Together with a similar argument for $h \uparrow 0$ this shows that F must be continuous. Hence the question becomes this: If F is continuous and f is integrable, and if (31.2) holds almost everywhere, does (31.1) follow? The answer, strangely enough, is still no: In Example 31.1 there is constructed a continuous, strictly increasing F for which $F'(x) = 0$ except on a set of Lebesgue measure 0, and (31.1) is of course impossible if f vanishes almost everywhere and F is strictly increasing. This leads to the problem of characterizing those F for which (31.1) does follow if (31.2) holds outside a set of Lebesgue measure 0 and f is integrable. In other words, which functions are the integrals of their (almost everywhere) derivatives? Theorem 31.8 gives the characterization.

It is possible to extend part (ii) of Theorem 31.1 in a different direction. Suppose that (31.2) holds for every x, not just almost everywhere. In Example 17.4 there was given a function F, everywhere differentiable, whose derivative f is not integrable, and in this case the right side of (31.1) has no meaning. If, however, (31.2) holds for every x, and if f is integrable, then (31.1) does hold for all x. For most purposes of probability theory, it is natural to impose conditions only almost everywhere, and so this theorem will not be proved here.[†]

The program then is first to show that (31.1) for integrable f implies that (31.2) holds almost everywhere, and second to characterize those F for which the reverse implication is valid. For the most part, f will be nonnegative and F will be nondecreasing. This is the case of greatest interest for probability theory; F can be regarded as a distribution function and f as a density.

[†]For a proof, see RUDIN$_2$, p 179

In Chapters 4 and 5 many distribution functions F were either shown to have a density f with respect to Lebesgue measure or were assumed to have one, but such F's were never intrinsically characterized, as they will be in this section.

Derivatives of Integrals

The first step is to show that a nondecreasing function has a derivative almost everywhere. This requires two preliminary results. Let λ denote Lebesgue measure.

Lemma 1. *Let A be a bounded linear Borel set, and let $\mathscr{I}$ be a collection of open intervals covering A. Then $\mathscr{I}$ contains a finite, disjoint subcollection $I_1, \ldots, I_k$ for which $\sum_{i=1}^{k} \lambda(I_i) \geq \lambda(A)/6$.*

PROOF. By regularity (Theorem 12.3) A contains a compact subset K satisfying $\lambda(K) \geq \lambda(A)/2$. Choose in $\mathscr{I}$ a finite subcollection $\mathscr{I}_0$ covering K. Let I_1 be an interval in $\mathscr{I}_0$ of maximal length; discard from $\mathscr{I}_0$ the interval I_1 and all the others that intersect I_1. Among the intervals remaining in $\mathscr{I}_0$, let I_2 be one of maximal length; discard I_2 and all intervals that intersect it. Continue this way until $\mathscr{I}_0$ is exhausted. The I_i are disjoint. Let J_i be the interval with the same midpoint as I_i and three times the length. If I is an interval in $\mathscr{I}_0$ that is cast out because it meets I_i, then $I \subset J_i$. Thus each discarded interval is contained in one of the J_i, and so the J_i cover K. Hence $\sum \lambda(I_i) = \sum \lambda(J_i)/3 \geq \lambda(K)/3 \geq \lambda(A)/6$. ∎

If

$$(31.4) \qquad \Delta: a = a_0 < a_1 < \cdots < a_k = b$$

is a partition of an interval $[a, b]$ and F is a function over $[a, b]$, let

$$(31.5) \qquad \|F\|_\Delta = \sum_{i=1}^{k} |F(a_i) - F(a_{i-1})|.$$

Lemma 2. *Consider a partition (31.4) and a nonnegative θ. If*

$$(31.6) \qquad F(a) \leq F(b),$$

and if

$$(31.7) \qquad \frac{F(a_i) - F(a_{i-1})}{a_i - a_{i-1}} \leq -\theta$$

for a set of intervals $[a_{i-1}, a_i]$ of total length d, then

$$\|F\|_\Delta \geq |F(b) - F(a)| + 2\theta d.$$

This also holds if the inequalities in (31.6) and (31.7) are reversed and $-\theta$ is replaced by θ in the latter.

PROOF. The figure shows the case where $k = 2$ and the left-hand interval satisfies (31.7). Here F falls at least θd over $[a, a + d]$, rises the same amount over $[a + d, u]$, and then rises $F(b) - F(a)$ over $[u, b]$.
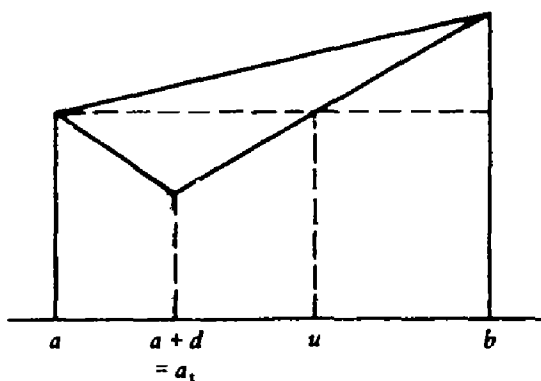
For the general case, let Σ' denote summation over those i satisfying (31.7) and let Σ'' denote summation over the remaining i ($1 \leq i \leq k$). Then

$$\|F\|_\Delta = \sum{}'(F(a_{i-1}) - F(a_i)) + \sum{}''|F(a_i) - F(a_{i-1})|$$

$$\geq \sum{}'(F(a_{i-1}) - F(a_i)) + \left|\sum{}''(F(a_i) - F(a_{i-1}))\right|$$

$$= \sum{}'(F(a_{i-1}) - F(a_i)) + \left|(F(b) - F(a)) + \sum{}'(F(a_{i-1}) - F(a_i))\right|.$$

As all the differences in this last expression are nonnegative, the absolute-value bars can be suppressed; therefore,

$$\|F\|_\Delta \geq F(b) - F(a) + 2\sum{}'(F(a_{i-1}) - F(a_i))$$

$$\geq F(b) - F(a) + 2\theta \sum{}'(a_i - a_{i-1}).$$ ∎

A function F has at each x four *derivates*, the upper and lower right derivatives

$$D^F(x) = \limsup_{h \downarrow 0} \frac{F(x + h) - F(x)}{h},$$

$$D_F(x) = \liminf_{h \downarrow 0} \frac{F(x + h) - F(x)}{h},$$

and the upper and lower left derivatives

$${}^F\!D(x) = \limsup_{h \downarrow 0} \frac{F(x) - F(x - h)}{h},$$

$${}_F\!D(x) = \liminf_{h \downarrow 0} \frac{F(x) - F(x - h)}{h}.$$

There is a derivative at x if and only if these four quantities have a common value. Suppose that F has finite derivative $F'(x)$ at x. If $u \le x \le v$, then

$$\left| \frac{F(v) - F(u)}{v - u} - F'(x) \right| \le \frac{v - x}{v - u} \left| \frac{F(v) - F(x)}{v - x} - F'(x) \right|$$
$$+ \frac{x - u}{v - u} \left| \frac{F(x) - F(u)}{x - u} - F'(x) \right|.$$

Therefore,

(31.8) $$\frac{F(v) - F(u)}{v - u} \to F'(x)$$

as $u \uparrow x$ and $v \downarrow x$; that is to say, for each ϵ there is a δ such that $u \le x \le v$ and $0 < v - u < \delta$ together imply that the quantities on either side of the arrow differ by less than ϵ.

Suppose that F is measurable and that it is continuous except possibly at countably many points. This will be true if F is nondecreasing or is the difference of two nondecreasing functions. Let M be a countable, dense set containing all the discontinuity points of F; let $r_n(x)$ be the smallest number of the form k/n exceeding x. Then

$$D^F(x) = \lim_{n \to \infty} \sup_{\substack{x < y < r_n(x) \\ y \in M}} \frac{F(y) - F(x)}{y - x};$$

the function inside the limit is measurable because the x-set where it exceeds α is

$$\bigcup_{y \in M} \left[x : x < y < r_n(x), F(y) - F(x) > \alpha(y - x) \right].$$

Thus $D^F(x)$ is measurable, as are the other three derivates. This does not exclude infinite values. The set where the four derivates have a common finite value F' is therefore a Borel set. In the following theorem, set $F' = 0$ (say) outside this set; F' is then a Borel function.

Theorem 31.2. *A nondecreasing function F is differentiable almost everywhere, the derivative F' is nonnegative, and*

(31.9) $$\int_a^b F'(t) \, dt \le F(b) - F(a)$$

for all a and b.

This and the following theorems can also be formulated for functions over an interval.

PROOF. If it can be shown that

$$(31.10) \qquad D^F(x) \leq {}_F D(x)$$

except on a set of Lebesgue measure 0, then by the same result applied to $G(x) = -F(-x)$ it will follow that ${}^F D(x) = D^G(-x) \leq {}_G D(-x) = D_F(x)$ almost everywhere. This will imply that $D_F(x) \leq D^F(x) \leq {}_F D(x) \leq {}^F D(x) \leq D_F(x)$ almost everywhere, since the first and third of these inequalities are obvious, and so, outside a set of Lebesgue measure 0, F will have a derivative, possibly infinite. Since F is nondecreasing, F' must be nonnegative, and once (31.9) is proved, it will follow that F' is finite almost everywhere.

If (31.10) is violated for a particular x, then for some pair α, β of rationals satisfying $\alpha < \beta$, x will lie in the set $A_{\alpha\beta} = [x: {}_F D(x) < \alpha < \beta < D^F(x)]$. Since there are only countably many of these sets, (31.10) will hold outside a set of Lebesgue measure 0 if $\lambda(A_{\alpha\beta}) = 0$ for all α and β.

Put $G(x) = F(x) - \frac{1}{2}(\alpha + \beta)x$ and $\theta = \frac{1}{2}(\beta - \alpha)$. Since differentiation is linear, $A_{\alpha\beta} = B_\theta = [x: {}_G D(x) < -\theta < \theta < D^G(x)]$. Since F and G have only countably many discontinuities, it suffices to prove that $\lambda(C_\theta) = 0$, where C_θ is the set of points in B_θ that are continuity points of G. Consider an interval (a, b), and suppose for the moment that $G(a) \leq G(b)$. For each x in C_θ satisfying $a < x < b$, from ${}_G D(x) < -\theta$ it follows that there exists an open interval (a_x, b_x) for which $x \in (a_x, b_x) \subset (a, b)$ and

$$(31.11) \qquad \frac{G(b_x) - G(a_x)}{b_x - a_x} < -\theta.$$

There exists by Lemma 1 a finite, disjoint collection (a_{x_i}, b_{x_i}) of these intervals of total length $\sum(b_{x_i} - a_{x_i}) \geq \lambda((a, b) \cap C_\theta)/6$. Let Δ be the partition (31.4) of $[a, b]$ with the points a_{x_i} and b_{x_i} in the role of the $a_1, \ldots, a_{k-1}$. By Lemma 2,

$$(31.12) \qquad \|G\|_\Delta \geq |G(b) - G(a)| + \tfrac{1}{3}\theta\lambda((a, b) \cap C_\theta).$$

If instead of $G(a) \leq G(b)$ the reverse inequality holds, choose a_x and b_x so that the ratio in (31.11) exceeds θ, which is possible because $D^G(x) > \theta$ for $x \in C_\theta$. Again (31.12) follows.

In each interval $[a, b]$ there is thus a partition (31.4) satisfying (31.12). Apply this to each interval $[a_{i-1}, a_i]$ in the partition. This gives a partition Δ_1 that refines Δ, and adding the corresponding inequalities (31.12) leads to

$$\|G\|_{\Delta_1} \geq \|G\|_\Delta + \tfrac{1}{3}\theta\lambda((a, b) \cap C_\theta).$$

Continuing leads to a sequence of successively finer partitions Δ_n such that

$$(31.13) \qquad \|G\|_{\Delta_n} \geq n\frac{\theta}{3}\lambda((a, b) \cap C_\theta).$$

Now $\|G\|_\Delta$ is bounded by $|F(b) - F(a)| + \frac{1}{2}|\alpha + \beta|(b - a)$ because F is monotonic. Thus (31.13) is impossible unless $\lambda((a, b) \cap C_\theta) = 0$. Since (a, b) can be any interval, $\lambda(C_\theta) = 0$. This proves (31.10) and establishes the differentiability of F almost everywhere.

It remains to prove (31.9). Let

$$(31.14) \qquad f_n(x) = \frac{F(x + n^{-1}) - F(x)}{n^{-1}}.$$

Now f_n is nonnegative, and by what has been shown, $f_n(x) \to F'(x)$ except on a set of Lebesgue measure 0. By Fatou's lemma and the fact that F is nondecreasing,

$$\int_a^b F'(x)\, dx \le \liminf_n \int_a^b f_n(x)\, dx$$

$$= \liminf_n \left[n \int_b^{b+n^{-1}} F(x)\, dx - n \int_a^{a+n^{-1}} F(x)\, dx \right]$$

$$\le \liminf_n \left[F(b + n^{-1}) - F(a) \right] = F(b+) - F(a).$$

Replacing b by $b - \epsilon$ and letting $\epsilon \to 0$ gives (31.9). ∎

Theorem 31.3. *If f is nonnegative and integrable, and if $F(x) = \int_{-\infty}^x f(t)\, dt$, then $F'(x) = f(x)$ except on a set of Lebesgue measure 0.*

Since f is nonnegative, F is nondecreasing and hence by Theorem 31.2 is differentiable almost everywhere. The problem is to show that the derivative F' coincides with f almost everywhere.

PROOF FOR BOUNDED f. Suppose first that f is bounded by M. Define f_n by (31.14). Then $f_n(x) = n\int_x^{x+n^{-1}} f(t)\, dt$ is bounded by M and converges almost everywhere to $F'(x)$, so that the bounded convergence theorem gives

$$\int_a^b F'(x)\, dx = \lim_n \int_a^b f_n(x)\, dx$$

$$= \lim_n \left[n \int_b^{b+n^{-1}} F(x)\, dx - n \int_a^{a+n^{-1}} F(x)\, dx \right].$$

Since F is continuous (see (31.3)), this last limit is $F(b) - F(a) = \int_a^b f(x)\, dx$.

Thus $\int_A F'(x)\, dx = \int_A f(x)\, dx$ for bounded intervals $A = (a, b]$. Since these form a π-system, it follows (Theorem 16.10(iii)) that $F' = f$ almost everywhere. ∎

PROOF FOR INTEGRABLE f. Apply the result for bounded functions to f truncated at n: If $h_n(x)$ is $f(x)$ or n as $f(x) \le n$ or $f(x) > n$, then $H_n(x) = \int_{-\infty}^x h_n(t)\,dt$ differentiates almost everywhere to $h_n(x)$ by the case already treated. Now $F(x) = H_n(x) + \int_{-\infty}^x (f(t) - h_n(t))\,dt$; the integral here is nondecreasing because the integrand is nonnegative, and it follows by Theorem 31.2 that it has almost everywhere a nonnegative derivative. Since differentiation is linear, $F'(x) \ge H_n'(x) = h_n(x)$ almost everywhere. As n was arbitrary, $F'(x) \ge f(x)$ almost everywhere, and so $\int_a^b F'(x)\,dx \ge \int_a^b f(x)\,dx = F(b) - F(a)$. But the reverse inequality is a consequence of (31.9). Therefore, $\int_a^b (F'(x) - f(x))\,dx = 0$, and as before $F' = f$ except on a set of Lebesgue measure 0. ∎

Singular Functions

If $f(x)$ is nonnegative and integrable, differentiating its indefinite integral $\int_{-\infty}^x f(t)\,dt$ leads back to $f(x)$ except perhaps on a set of Lebesgue measure 0. That is the content of Theorem 31.3. The converse question is this: If $F(x)$ is nondecreasing and hence has almost everywhere a derivative $F'(x)$, does integrating $F'(x)$ lead back to $F(x)$? As stated before, the answer turns out to be no even if $F(x)$ is assumed continuous:

Example 31.1. Let $X_1, X_2, \dots$ be independent, identically distributed random variables such that $P[X_n = 0] = p_0$ and $P[X_n = 1] = p_1 = 1 - p_0$, and let $X = \sum_{n=1}^\infty X_n 2^{-n}$. Let $F(x) = P[X \le x]$ be the distribution function of X. For an arbitrary sequence $u_1, u_2, \dots$ of 0's and 1's, $P[X_n = u_n, n = 1, 2, \dots] = \lim_n p_{u_1} \cdots p_{u_n} = 0$; since x can have at most two dyadic expansions $x = \sum_n u_n 2^{-n}$, $P[X = x] = 0$. Thus F is everywhere continuous. Of course, $F(0) = 0$ and $F(1) = 1$. For $0 \le k < 2^n$, $k 2^{-n}$ has the form $\sum_{i=1}^n u_i 2^{-i}$ for some n-tuple $(u_1, \dots, u_n)$ of 0's and 1's. Since F is continuous,

$$(31.15) \qquad F\!\left(\frac{k+1}{2^n}\right) - F\!\left(\frac{k}{2^n}\right) = P\!\left[\frac{k}{2^n} < X < \frac{k+1}{2^n}\right]$$

$$= P[X_i = u_i, i \le n] = p_{u_1} \cdots p_{u_n}.$$

This shows that F is strictly increasing over the unit interval.

If $p_0 = p_1 = \frac{1}{2}$, the right side of (31.15) is 2^{-n}, and a passage to the limit shows that $F(x) = x$ for $0 \le x \le 1$. Assume, however, that $p_0 \ne p_1$. It will be shown that $F'(x) = 0$ except on a set of Lebesgue measure 0 in this case. Obviously the derivative is 0 outside the unit interval, and by Theorem 31.2 it exists almost everywhere inside it. Suppose then that $0 < x < 1$ and that F has a derivative $F'(x)$ at x. It will be shown that $F'(x) = 0$.

For each n choose k_n so that x lies in the interval $I_n = (k_n 2^{-n}, (k_n + 1)2^{-n}]$; I_n is that dyadic interval of rank n that contains x. By (31.8),

$$\frac{P[X \in I_n]}{2^{-n}} = \frac{F((k_n + 1)2^{-n}) - F(k_n 2^{-n})}{2^{-n}} \to F'(x).$$
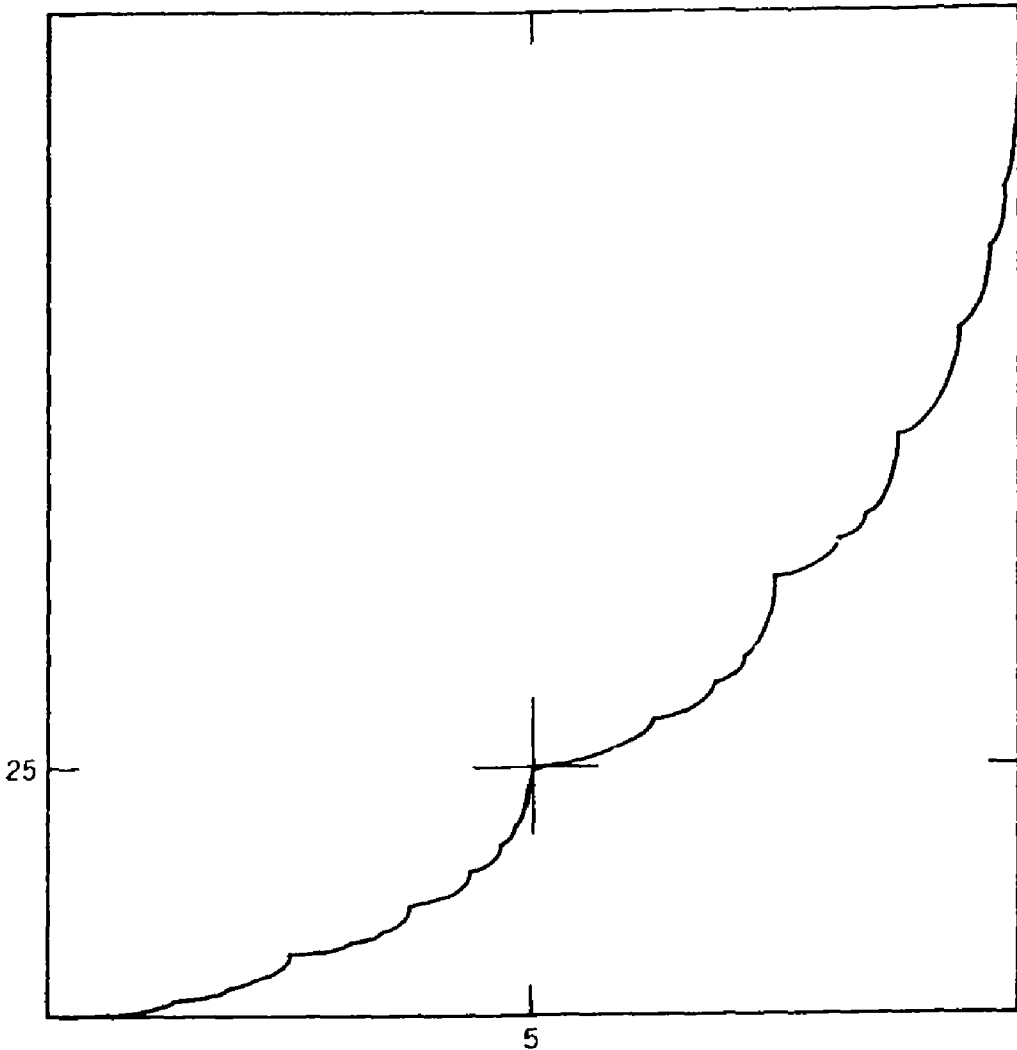
Graph of $F(x)$ for $p_0 = .25$, $p_1 = .75$. Because of the recursion (31.17), the part of the graph over $[0, .5]$ and the part over $[.5, 1]$ are identical, apart from changes in scale, with the whole graph Each segment of the curve therefore contains scaled copies of the whole, the extreme irregularity this implies is obscured by the fact that the accuracy is only to within the width of the printed line.

If $F'(x)$ is distinct from 0, the ratio of two successive terms here must go to 1, so that

$$(31.16) \qquad \frac{P[X \in I_{n+1}]}{P[X \in I_n]} \to \frac{1}{2}.$$

If I_n consists of the reals with nonterminating base-2 expansions beginning with the digits $u_1, \ldots, u_n$, then $P[X \in I_n] = p_{u_1} \cdots p_{u_n}$ by (31.15). But I_{n+1} must for some u_{n+1} consist of the reals beginning $u_1, \ldots, u_n, u_{n+1}$ (u_{n+1} is 1 or 0 according as x lies to the right of the midpoint of I_n or not). Thus $P[X \in I_{n+1}]/P[X \in I_n] = p_{u_{n+1}}$ is either p_0 or p_1, and (31.16) is possible only if $p_0 = p_1$, which was excluded by hypothesis.

Thus F is continuous and strictly increasing over $[0, 1]$, but $F'(x) = 0$ except on a set of Lebesgue measure 0. For $0 \le x \le \frac{1}{2}$ independence gives

$F(x) = P[X_1 = 0, \sum_{n=2}^{\infty} X_n 2^{-n+1} \leq 2x] = p_0 F(2x)$. Similarly, $F(x) - p_0 = p_1 F(2x - 1)$ for $\frac{1}{2} \leq x \leq 1$. Thus

$$(31.17) \qquad F(x) = \begin{cases} p_0 F(2x) & \text{if } 0 \leq x \leq \frac{1}{2}, \\ p_0 + p_1 F(2x - 1) & \text{if } \frac{1}{2} \leq x \leq 1. \end{cases}$$

In Section 7, $F(x)$ (there denoted $Q(x)$) entered as the probability of success at bold play; see (7.30) and (7.33). ∎

A function is *singular* if it has derivative 0 except on a set of Lebesgue measure 0. Of course, a step function constant over intervals is singular. What is remarkable (indeed, singular) about the function in the preceding example is that it is continuous and strictly increasing but nonetheless has derivative 0 except on a set of Lebesgue measure 0. Note that there is strict inequality in (31.9) for this F.

Further properties of nondecreasing functions can be discovered through a study of the measures they generate. Assume from now on that F is nondecreasing, that F is continuous from the right (this is only a normalization), and that $0 = \lim_{x \to -\infty} F(x) \leq \lim_{x \to +\infty} F(x) = m < \infty$. Call such an F a *distribution function*, even though m need not be 1. By Theorem 12.4 there exists a unique measure μ on the Borel sets of the line for which

$$(31.18) \qquad \mu(a, b] = F(b) - F(a).$$

Of course, $\mu(R^1) = m$ is finite.

The larger F' is, the larger μ is:

Theorem 31.4. *Suppose that F and μ are related by (31.18) and that $F'(x)$ exists throughout a Borel set A.*

(i) *If $F'(x) \leq \alpha$ for $x \in A$, then $\mu(A) \leq \alpha \lambda(A)$.*
(ii) *If $F'(x) \geq \alpha$ for $x \in A$, then $\mu(A) \geq \alpha \lambda(A)$.*

PROOF. It is no restriction to assume A bounded. Fix ϵ for the moment. Let E be a countable, dense set, and let $A_n = \bigcap (A \cap I)$, where the intersection extends over the intervals $I = (u, v]$ for which $u, v \in E$, $0 < \lambda(I) < n^{-1}$, and

$$(31.19) \qquad \mu(I) < (\alpha + \epsilon)\lambda(I).$$

Then A_n is a Borel set and (see (31.8)) $A_n \uparrow A$ under the hypothesis of (i). By Theorem 11.4 there exist disjoint intervals I_{nk} (open on the left, closed on

the right) such that $A_n \subset \bigcup_k I_{nk}$ and

$$(31.20) \qquad \sum_k \lambda(I_{nk}) < \lambda(A_n) + \epsilon.$$

It is no restriction to assume that each I_{nk} has endpoints in E, meets A_n, and satisfies $\lambda(I_{nk}) < n^{-1}$. Then (31.19) applies to each I_{nk}, and hence

$$\mu(A_n) \le \sum_k \mu(I_{nk}) \le (\alpha + \epsilon) \sum_k \lambda(I_{nk}) \le (\alpha + \epsilon)(\lambda(A_n) + \epsilon).$$

In the extreme terms here let $n \to \infty$ and then $\epsilon \to 0$; (i) follows.

To prove (ii), let the countable, dense set E contain all the discontinuity points of F, and use the same argument with $\mu(I) \ge (\alpha - \epsilon)\lambda(I)$ in place of (31.19) and $\sum_k \mu(I_{nk}) < \mu(A_n) + \epsilon$ in place of (31.20). Since E contains all the discontinuity points of F, it is again no restriction to assume that each I_{nk} has endpoints in E, meets A_n, and satisfies $\lambda(I_{nk}) < n^{-1}$. It follows that

$$\mu(A_n) + \epsilon > \sum_k \mu(I_{nk}) \ge (\alpha - \epsilon) \sum_k \lambda(I_{nk}) \ge (\alpha - \epsilon)\lambda(A_n).$$

Again let $n \to \infty$ and then $\epsilon \to 0$.   ∎

The measures μ and λ have *disjoint supports* if there exist Borel sets S_μ and S_λ such that

$$(31.21) \qquad \begin{cases} \mu(R^1 - S_\mu) = 0, \qquad \lambda(R^1 - S_\lambda) = 0, \\ S_\mu \cap S_\lambda = 0. \end{cases}$$

Theorem 31.5. *Suppose that F and μ are related by (31.18). A necessary and sufficient condition for μ and λ to have disjoint supports is that $F'(x) = 0$ except on a set of Lebesgue measure 0.*

PROOF. By Theorem 31.4, $\mu[x: |x| \le a, F'(x) \le \epsilon] \le 2a\epsilon$, and so (let $\epsilon \to 0$ and then $a \to \infty$) $\mu[x: F'(x) = 0] = 0$. If $F'(x) = 0$ outside a set of Lebesgue measure 0, then $S_\lambda = [x: F'(x) = 0]$ and $S_\mu = R^1 - S_\lambda$ satisfy (31.21).

Suppose that there exist S_μ and S_λ satisfying (31.21). By the other half of Theorem 31.4, $\epsilon\lambda[x: F'(x) \ge \epsilon] = \epsilon\lambda[x: x \in S_\lambda, F'(x) \ge \epsilon] \le \mu(S_\lambda) = 0$, and so (let $\epsilon \to 0$) $F'(x) = 0$ except on a set of Lebesgue measure 0.   ∎

Example 31.2. Suppose that μ is discrete, consisting of a mass m_k at each of countably many points x_k. Then $F(x) = \sum m_k$, the sum extending over the k for which $x_k \le x$. Certainly, μ and λ have disjoint supports, and so F' must vanish except on a set of Lebesgue measure 0. This is directly obvious if the x_k have no limit points, but not, for example, if they are dense.   ∎

Example 31.3. Consider again the distribution function F in Example 31.1. Here $\mu(A) = P[X \in A]$. Since F is singular, μ and λ have disjoint supports. This fact has an interesting direct probabilistic proof.

For x in the unit interval, let $d_1(x), d_2(x), \ldots$ be the digits in its nonterminating dyadic expansion, as in Section 1. If $(k2^{-n}, (k+1)2^{-n}]$ is the dyadic interval of rank n consisting of the reals whose expansions begin with the digits $u_1, \ldots, u_n$, then, by (31.15),

$$(31.22) \qquad \mu\left(\frac{k}{2^n}, \frac{k+1}{2^n}\right] = \mu[x: d_i(x) = u_i, i \le n] = p_{u_1} \cdots p_{u_n}.$$

If the unit interval is regarded as a probability space under the measure μ, then the $d_i(x)$ become random variables, and (31.22) says that these random variables are independent and identically distributed and $\mu[x: d_i(x) = 0] = p_0$, $\mu[x: d_i(x) = 1] = p_1$.

Since these random variables have expected value p_1, the strong law of large numbers implies that their averages go to p_1 with probability 1:

$$(31.23) \qquad \mu\left[x \in (0,1]: \lim_n \frac{1}{n} \sum_{i=1}^n d_i(x) = p_1\right] = 1.$$

On the other hand, by the normal number theorem,

$$(31.24) \qquad \lambda\left[x \in (0,1]: \lim_n \frac{1}{n} \sum_{i=1}^n d_i(x) = \frac{1}{2}\right] = 1.$$

(Of course, (31.24) is just (31.23) for the special case $p_0 = p_1 = \frac{1}{2}$; in this case μ and λ coincide in the unit interval.) If $p_1 \ne \frac{1}{2}$, the sets in (31.23) and (31.24) are disjoint, so that μ and λ do have disjoint supports.

It was shown in Example 31.1 that if $F'(x)$ exists at all $(0 < x < 1)$, then it is 0. By part (i) of Theorem 31.4 the set where $F'(x)$ fails to exist therefore has μ-measure 1; in particular, this set is uncountable. ∎

In the singular case, according to Theorem 31.5, F' vanishes on a support of λ. It is natural to ask for the size of F' on a support of μ. If B is the x-set where F has a finite derivative, and if (31.21) holds, then by Theorem 31.4, $\mu[x \in B: F'(x) \le n] = \mu[x \in B \cap S_\mu: F'(x) \le n] \le n\lambda(S_\mu) = 0$, and hence $\mu(B) = 0$. The next theorem goes further.

Theorem 31.6. *Suppose that F and μ are related by (31.18) and that μ and λ have disjoint supports. Then, except for x in a set of μ-measure 0, $_FD(x) = \infty$.*

If μ has finite support, then clearly $_FD(x) = \infty$ if $\mu\{x\} > 0$, while $D_F(x) = 0$ for all x. Since F is continuous from the right, $_FD$ and D_F play different roles.[†]

PROOF. Let A_n be the set where $_FD(x) < n$. The problem is to prove that $\mu(A_n) = 0$, and by (31.21) it is enough to prove that $\mu(A_n \cap S_\mu) = 0$. Further, by Theorem 12.3 it is enough to prove that $\mu(K) = 0$ if K is a compact subset of $A_n \cap S_\mu$.

Fix ϵ. Since $\lambda(K) = 0$, there is an open G such that $K \subset G$ and $\lambda(G) < \epsilon$. If $x \in K$, then $x \in A_n$, and by the definition of $_FD$ and the right-continuity of F, there is an open interval I_x for which $x \in I_x \subset G$ and $\mu(I_x) < n\lambda(I_x)$. By compactness, K has a finite subcover $I_{x_1}, \ldots, I_{x_k}$. If some three of these have a nonempty intersection, one of them must be contained in the union of the other two. Such superfluous intervals can be removed from the subcover, and it is therefore possible to assume that no point of K lies in more than two of the I_{x_i}. But then

$$\mu(K) \leq \mu\left(\bigcup_i I_{x_i}\right) \leq \sum_i \mu(I_{x_i}) \leq n \sum_i \lambda(I_{x_i})$$

$$\leq 2n\lambda\left(\bigcup_i I_{x_i}\right) \leq 2n\lambda(G) \leq 2n\epsilon.$$

Since ϵ was arbitrary, $\lambda(K) = 0$, as required.   ∎

Example 31.4. Restrict the F of Examples 31.1 and 31.3 to $(0,1)$, and let g be the inverse. Thus F and g are continuous, strictly increasing mappings of $(0,1)$ onto itself. If $A = [x \in (0,1): F'(x) = 0]$, then $\lambda(A) = 1$, as shown in the examples, while $\mu(A) = 0$. Let H be a set in $(0,1)$ that is not a Lebesgue set. Since $H - A$ is contained in a set of Lebesgue measure 0, it is a Lebesgue set; hence $H_0 = H \cap A$ is not a Lebesgue set, since otherwise $H = H_0 \cup (H - A)$ would be a Lebesgue set. If $B = (0, x]$, then $\lambda g^{-1}(B) = \lambda(0, F(x)] = F(x) = \mu(B)$, and it follows that $\lambda g^{-1}(B) = \mu(B)$ for all Borel sets B. Since $g^{-1}H_0$ is a subset of $g^{-1}A$ and $\lambda(g^{-1}A) = \mu(A) = 0$, $g^{-1}H_0$ is a Lebesgue set. On the other hand, if $g^{-1}H_0$ were a Borel set, $H_0 = F^{-1}(g^{-1}H_0)$ would also be a Borel set. Thus $g^{-1}H_0$ provides an example of a Lebesgue set that is not a Borel set.[‡]   ∎

Integrals of Derivatives

Return now to the problem of extending part (ii) of Theorem 31.1, to the problem of characterizing those distribution functions F for which F' integrates back to F:

$$(31.25) \qquad\qquad F(x) = \int_{-\infty}^x F'(t)\, dt.$$

[†]See Problem 31.8
[‡]For a different argument, see Problem 3.14.

The first step is easy: If (31.25) holds, then F has the form

$$(31.26) \qquad\qquad F(x) = \int_{-\infty}^{x} f(t)\, dt$$

for a nonnegative, integrable f (a density), namely $f = F'$. On the other hand, (31.26) implies by Theorem 31.3 that $F' = f$ outside a set of Lebesgue measure 0, whence (31.25) follows. Thus (31.25) holds if and only if F has the form (31.26) for some f, and the problem is to characterize functions of this form. The function of Example 31.1 is not among them.

As observed earlier (see (31.3)), an F of the form (31.26) with f integrable is continuous. It has a still stronger property: For each ϵ there exists a δ such that

$$(31.27) \qquad\qquad \int_A f(x)\, dx < \epsilon \qquad \text{if } \lambda(A) < \delta.$$

Indeed, if $A_n = [x: f(x) > n]$, then $A_n \downarrow \varnothing$, and since f is integrable, the dominated convergence theorem implies that $\int_{A_n} f(x)\, dx < \epsilon/2$ for large n. Fix such an n and take $\delta = \epsilon/2n$. If $\lambda(A) < \delta$, then $\int_A f(x)\, dx \le \int_{A - A_n} f(x)\, dx + \int_{A_n} f(x)\, dx \le n\lambda(A) + \epsilon/2 < \epsilon$.

If F is given by (31.26), then $F(b) - F(a) = \int_a^b f(x)\, dx$, and (31.27) has this consequence: For every ϵ there exists a δ such that for each finite collection $[a_i, b_i]$, $i = 1, \dots, k$, of nonoverlapping† intervals,

$$(31.28) \qquad \sum_{i=1}^{k} |F(b_i) - F(a_i)| < \epsilon \qquad \text{if } \sum_{i=1}^{k} (b_i - a_i) < \delta.$$

A function F with this property is said to be *absolutely continuous*.‡ A function of the form (31.26) (f integrable) is thus absolutely continuous.

A continuous distribution function is uniformly continuous, and so for every ϵ there is a δ such that the implication in (31.28) holds provided that $k = 1$. The definition of absolute continuity requires this to hold whatever k may be, which puts severe restrictions on F. Absolute continuity of F can be characterized in terms of the measure μ:

Theorem 31.7. *Suppose that F and μ are related by (31.18). Then F is absolutely continuous in the sense of (31.28) if and only if $\mu(A) = 0$ for every A for which $\lambda(A) = 0$.*

---

†Intervals are nonoverlapping if their interiors are disjoint. In this definition it is immaterial whether the intervals are regarded as closed or open or half-open, since this has no effect on (31.28).

‡The definition applies to all functions, not just to distribution functions. If F is a distribution function as in the present discussion, the absolute-value bars in (31.28) are unnecessary

PROOF. Suppose that F is absolutely continuous and that $\lambda(A) = 0$. Given ϵ, choose δ so that (31.28) holds. There exists a countable disjoint union $B = \bigcup_k I_k$ of intervals such that $A \subset B$ and $\lambda(B) < \delta$. By (31.28) it follows that $\mu(\bigcup_{k=1}^n I_k) \le \epsilon$ for each n and hence that $\mu(A) \le \mu(B) \le \epsilon$. Since ϵ was arbitrary, $\mu(A) = 0$.

If F is not absolutely continuous, then there exists an ϵ such that for every δ some finite disjoint union A of intervals satisfies $\lambda(A) < \delta$ and $\mu(A) \ge \epsilon$. Choose A_n so that $\lambda(A_n) < n^{-2}$ and $\mu(A_n) \ge \epsilon$. Then $\lambda(\limsup_n A_n) = 0$ by the first Borel–Cantelli lemma (Theorem 4.3, the proof of which does not require P to be a probability measure or even finite). On the other hand, $\mu(\limsup_n A_n) \ge \epsilon > 0$ by Theorem 4.1 (the proof of which applies because μ is assumed finite). ∎

This result leads to a characterization of indefinite integrals.

Theorem 31.8. *A distribution function $F(x)$ has the form $\int_{-\infty}^x f(t)\,dt$ for an integrable f if and only if it is absolutely continuous in the sense of (31.28).*

PROOF. That an F of the form (31.26) is absolutely continuous was proved in the argument leading to the definition (31.28). For another proof, apply Theorem 31.7: if F has this form, then $\lambda(A) = 0$ implies that $\mu(A) = \int_A f(t)\,dt = 0$.

To go the other way, define for any distribution function F

$$(31.29) \qquad F_{ac}(x) = \int_{-\infty}^x F'(t)\,dt$$

and

$$(31.30) \qquad F_s(x) = F(x) - F_{ac}(x).$$

Then F_s is right-continuous, and by (31.9) it is both nonnegative and nondecreasing. Since F_{ac} comes form a density, it is absolutely continuous. By Theorem 31.3, $F'_{ac} = F'$ and hence $F'_s = 0$ except on a set of Lebesgue measure 0. Thus F has a decomposition

$$(31.31) \qquad F(x) = F_{ac}(x) + F_s(x),$$

where F_{ac} has a density and hence is absolutely continuous and F_s is singular. This is called the *Lebesgue decomposition*.

Suppose that F is absolutely continuous. Then F_s of (31.30) must, as the difference of absolutely continuous functions, be absolutely continuous itself. If it can be shown that F_s is identically 0, it will follow that $F = F_{ac}$ has the required form. It thus suffices to show that a distribution function that is both absolutely continuous and singular must vanish.

If a distribution function F is singular, then by Theorem 31.5 there are disjoint supports S_μ and S_λ. But if F is also absolutely continuous, then from $\lambda(S_\mu) = 0$ it follows by Theorem 31.7 that $\mu(S_\mu) = 0$. But then $\mu(R^1) = 0$, and so $F(x) \equiv 0$. ■

This theorem identifies the distribution functions that are integrals of their derivatives as the absolutely continuous functions. Theorem 31.7, on the other hand, characterizes absolute continuity in a way that extends to spaces Ω without the geometric structure of the line necessary to a treatment involving distribution functions and ordinary derivatives.[†] The extension is studied in Section 32.

Functions of Bounded Variation

The remainder of this section briefly sketches the extension of the preceding theory to functions that are not monotone. The results are for simplicity given only for a finite interval $[a, b]$ and for functions F on $[a, b]$ satisfying $F(a) = 0$.

If $F(x) = \int_a^x f(t)\, dt$ is an indefinite integral, where f is integrable but not necessarily nonnegative, then $F(x) = \int_a^x f^+(t)\, dt - \int_a^x f^-(t)\, dt$ exhibits F as the difference of two nondecreasing functions. The problem of characterizing indefinite integrals thus leads to the preliminary problem of characterizing functions representable as a difference of nondecreasing functions.

Now F is said to be of *bounded variation* over $[a, b]$ if $\sup_\Delta \|F\|_\Delta$ is finite, where $\|F\|_\Delta$ is defined by (31.5) and Δ ranges over all partitions (31.4) of $[a, b]$. Clearly, a difference of nondecreasing functions is of bounded variation. But the converse holds as well: For every finite collection Γ of nonoverlapping intervals $[x_i, y_i]$ in $[a, b]$, put

$$P_\Gamma = \sum (F(y_i) - F(x_i))^+, \qquad N_\Gamma = \sum (F(y_i) - F(x_i))^-.$$

Now define

$$P(x) = \sup_\Gamma P_\Gamma, \qquad N(x) = \sup_\Gamma N_\Gamma,$$

where the suprema extend over partitions Γ of $[a, x]$. If F is of bounded variation, then $P(x)$ and $N(x)$ are finite. For each such Γ, $P_\Gamma = N_\Gamma + F(x)$. This gives the inequalities

$$P_\Gamma \le N(x) + F(x), \qquad P(x) \ge N_\Gamma + F(x),$$

which in turn lead to the inequalities

$$P(x) \le N(x) + F(x), \qquad P(x) \ge N(x) + F(x).$$

Thus

(31.32) $$F(x) = P(x) - N(x)$$

gives the required representation: *A function is the difference of two nondecreasing functions if and only if it is of bounded variation.*

[†]Theorems 31.3 and 31 8 do have geometric analogues in R^k; see RUDIN₂, Chapter 8.

If $T_\Gamma = P_\Gamma + N_\Gamma$, then $T_\Gamma = \Sigma |F(y_i) - F(x_i)|$. According to the definition (31.28), F is absolutely continuous if for every ϵ there exists a δ such that $T_\Gamma < \epsilon$ whenever the intervals in the collection Γ have total length less than δ. If F is absolutely continuous, take the δ corresponding to an ϵ of 1 and decompose $[a, b]$ into a finite number, say n, of subintervals $[u_{j-1}, u_j]$ of lengths less than δ. Any partition Δ of $[a, b]$ can by the insertion of the u_j be split into n sets of intervals each of total length less than δ, and it follows[t] that $\|F\|_\Delta \le n$. Therefore, an absolutely continuous function is necessarily of bounded variation.

An absolutely continuous F thus has a representation (31.32). It follows by the definitions that $P(y) - P(x)$ is at most $\sup_\Gamma T_\Gamma$, where Γ ranges over the partitions of $[x, y]$. If $[x_i, y_i]$ are nonoverlapping intervals, then $\Sigma(P(y_i) - P(x_i))$ is at most $\sup_\Gamma T_\Gamma$, where now Γ ranges over the collections of intervals that partition *each* of the $[x_i, y_i]$. Therefore, if F is absolutely continuous, there exists for each ϵ a δ such that $\Sigma(y_i - x_i) < \delta$ implies that $\Sigma(P(y_i) - P(x_i)) < \epsilon$. In other words, P is absolutely continuous. Similarly, N is absolutely continuous.

Thus an absolutely continuous F is the difference of two nondecreasing absolutely continuous functions. By Theorem 31.8, each of these is an indefinite integral, which implies that F is an indefinite integral as well: *For an F on $[a, b]$ satisfying $F(a) = 0$, absolute continuity is a necessary and sufficient condition for F to be an indefinite integral—to have the form $F(x) = \int_a^x f(t)\, dt$ for an integrable f.*

PROBLEMS

31.1. Extend Examples 31.1 and 31.3: Let $p_0, \ldots, p_{r-1}$ be nonnegative numbers adding to 1, where $r \ge 2$; suppose there is no i such that $p_i = 1$. Let X_1, $X_2, \ldots$ be independent, identically distributed random variables such that $P[X_n = i] = p_i$, $0 \le i < r$, and put $X = \sum_{n=1}^\infty X_n r^{-n}$. Let F be the distribution function of X. Show that F is continuous. Show that F is strictly increasing over the unit interval if and only if all the p_i are strictly positive. Show that $F(x) \equiv x$ for $0 \le x \le 1$ if $p_i \equiv r^{-1}$ and that otherwise F is singular; prove singularity by extending the arguments both of Example 31.1 and of Example 31.3. What is the analogue of (31.17)?

31.2. ↑   In Problem 31.1 take $r = 3$ and $p_0 = p_2 = \frac{1}{2}$, $p_1 = 0$. The corresponding F is called the *Cantor function*. The complement in $[0, 1]$ of the Cantor set (see Problems 1.5 and 3.16) consists of the middle third $(\frac{1}{3}, \frac{2}{3})$, the middle thirds $(\frac{1}{9}, \frac{2}{9})$ and $(\frac{7}{9}, \frac{8}{9})$, and so on. Show that F is $\frac{1}{2}$ on the first of these intervals, $\frac{1}{4}$ on the second, $\frac{3}{4}$ on the third, and so on. Show by direct argument that $F' = 0$ except on a set of Lebesgue measure 0.

31.3. A real function f of a real variable is a *Lebesgue function* if $[x: f(x) \le \alpha]$ is a Lebesgue set for each α.

(a) Show that, if f_1 is a Borel function and f_2 is a Lebesgue function, then the composition $f_1 f_2$ is a Lebesgue function.

(b) Show that there exists a Lebesgue function f_1 and a Lebesgue (even Borel, even continuous) function f_2 such that $f_1 f_2$ is not a Lebesgue function. *Hint:* Use Example 31.4.

[t]This uses the fact that $\|F\|_\Delta$ cannot decrease under passage to a finer partition.

31.4. ↑ An arbitrary function f on $(0,1]$ can be represented as a composition of a Lebesgue function f_1 and a Borel function f_2. For x in $(0,1]$, let $d_n(x)$ be the nth digit in its nonterminating dyadic expansion, and define $f_2(x) = \sum_{n=1}^{\infty} 2d_n(x)/3^n$. Show that f_2 is increasing and that $f_2(0,1]$ is contained in the Cantor set. Take $f_1(x)$ to be $f(f_2^{-1}(x))$ if $x \in f_2(0,1]$ and 0 if $x \in (0,1] - f_2(0,1]$. Now show that $f = f_1 f_2$.

31.5. Let $r_1, r_2, \ldots$ be an enumeration of the rationals in $(0,1)$ and put $F(x) = \sum_{k: r_k \leq x} 2^{-k}$. Define φ by (14.5) and prove that it is continuous and singular.

31.6. Suppose that μ and F are related by (31.18). If F is not absolutely continuous, then $\mu(A) > 0$ for some set A of Lebesgue measure 0. It is an interesting fact, however, that almost all translates of A must have μ-measure 0. From Fubini's theorem and the fact that λ is invariant under translation and reflection through 0, show that, if $\lambda(A) = 0$ and μ is σ-finite, then $\mu(A + x) = 0$ for x outside a set of Lebesgue measure 0.

31.7. 17.4 31.6↑ Show that F is absolutely continuous if and only if for each Borel set A, $\mu(A + x)$ is continuous in x.

31.8. Let $F_*(x) = \lim_{\delta \to 0} \inf(F(v) - F(u))/(v - u)$, where the infimum extends over u and v such that $u < x < v$ and $v - u < \delta$. Define $F^*(x)$ as this limit with the infimum replaced by a supremum. Show that in Theorem 31.4, F' can be replaced by F^* in part (i) and by F_* in part (ii). Show that in Theorem 31.6, $_FD$ can be replaced by F_* (note that $F_*(x) \leq _FD(x)$).

31.9. *Lebesgue's density theorem.* A point x is a *density point* of a Borel set A if $\lambda((u,v] \cap A)/(v - u) \to 1$ as $u \uparrow x$ and $v \downarrow x$. From Theorems 31.2 and 31.4 deduce that almost all points of A are density points. Similarly, $\lambda((u,v] \cap A)/(v - u) \to 0$ almost everywhere on A^c.

31.10. Let $f: [a,b] \to R^k$ be an arc; $f(t) = (f_1(t), \ldots, f_k(t))$. Show that the arc is rectifiable if and only if each f_i is of bounded variation over $[a,b]$.

31.11. ↑ Suppose that F is continuous and nondecreasing and that $F(0) = 0$, $F(1) = 1$. Then $f(x) = (x, F(x))$ defines an arc $f: [0,1] \to R^2$. It is easy to see by monotonicity that the arc is rectifiable and that, in fact, its length satisfies $L(f) \leq 2$. It is also easy, given ϵ, to produce functions F for which $L(f) > 2 - \epsilon$. Show by the arguments in the proof of Theorem 31.4 that $L(f) = 2$ if F is singular.

31.12. Suppose that the characteristic function of F satisfies $\limsup_{t \to \infty} |\varphi(t)| = 1$. Show that F is singular. Compare the lattice case (Problem 26.1). *Hint:* Use the Lebesgue decomposition and the Riemann–Lebesgue theorem.

31.13. Suppose that $X_1, X_2, \ldots$ are independent and assume the values ± 1 with probability $\frac{1}{2}$ each, and let $X = \sum_{n=1}^{\infty} X_n/2^n$. Show that X is uniformly distributed over $[-1, +1]$. Calculate the characteristic functions of X and X_n and deduce (1.40). Conversely, establish (1.40) by trigonometry and conclude that X is uniformly distributed over $[-1, +1]$.

31.14. (a) Suppose that $X_1, X_2, \ldots$ are independent and assume the values 0 and 1 with probability $\frac{1}{2}$ each. Let F and G be the distribution functions of $\sum_{n=1}^{\infty} X_{2n-1}/2^{2n-1}$ and $\sum_{n=1}^{\infty} X_{2n}/2^{2n}$. Show that F and G are singular but that $F * G$ is absolutely continuous.

(b) Show that the convolution of an absolutely continuous distribution function with an arbitrary distribution function is absolutely continuous.

31.15. 31.2↑ Show that the Cantor function is the distribution function of $\sum_{n=1}^{\infty} X_n/3^n$, where the X_n are independent and assume the values 0 and 2 with probability $\frac{1}{2}$ each. Express its characteristic function as an infinite product.

31.16. Show for the F of Example 31.1 that $_F D(1) = \infty$ and $D^F(0) = 0$ if $p_0 < \frac{1}{2}$. From (31 17) deduce that $_F D(x) = \infty$ and $D^F(x) = 0$ for all dyadic rationals x. Analyze the case $p_0 > \frac{1}{2}$ and sketch the graph

31.17. 6.14↑ Let F be as in Example 31.1, and let μ be the corresponding probability measure on the unit interval. Let $d_n(x)$ be the nth digit in the nonterminating binary expansion of x, and let $s_n(x) = \sum_{k=1}^{n} d_k(x)$. If $I_n(x)$ is the dyadic interval of order n containing x, then

$$(31.33) \quad -\frac{1}{n} \log \mu(I_n(x)) = -\left(1 - \frac{s_n(x)}{n}\right) \log p_0 - \frac{s_n(x)}{n} \log p_1.$$

(a) Show that (31.33) converges on a set of μ-measure 1 to the entropy $h = -p_0 \log p_0 - p_1 \log p_1$. From the fact that this entropy is less than $\log 2$ if $p_0 \neq \frac{1}{2}$, deduce in this case that on a set of μ-measure 1, F does not have a finite derivative.

(b) Show that (31.33) converges to $-\frac{1}{2} \log p_0 - \frac{1}{2} \log p_1$ on a set of Lebesgue measure 1. If $p_0 \neq \frac{1}{2}$ this limit exceeds $\log 2$ (arithmetic versus geometric means), and so $\mu(I_n(x))/2^{-n} \to 0$ except on a set of Lebesgue measure 0. This does not prove that $F'(x)$ exists almost everywhere, but it does show that, except for x in a set of Lebesgue measure 0. if $F'(x)$ does exist, then it is 0.

(c) Show that, if (31.33) converges to l, then

$$(31.34) \qquad \lim_n \frac{\mu(I_n(x))}{(2^{-n})^\alpha} = \begin{cases} \infty & \text{if } \alpha > l/\log 2, \\ 0 & \text{if } \alpha < l/\log 2. \end{cases}$$

If (31.34) holds, then (roughly) F satisfies a Lipschitz condition[†] of (exact) order $l/\log 2$. Thus F satisfies a Lipschitz condition of order $h/\log 2$ on a set of μ-measure 1 and a Lipschitz condition of order $(-\frac{1}{2} \log p_0 - \frac{1}{2} \log p_1)/\log 2$ on a set of Lebesgue measure 1.

31.18. van der Waerden's continuous, nowhere differentiable function is $f(x) = \sum_{k=0}^{\infty} a_k(x)$, where $a_0(x)$ is the distance from x to the nearest integer and $a_k(x) = 2^{-k} a_0(2^k x)$. Show by the Weierstrass M-test that f is continuous. Use (31.8) and the ideas in Example 31.1 to show that f is nowhere differentiable.

[†]A Lipschitz condition of order α holds at x if $F(x + h) - F(x) = O(|h|^\alpha)$ as $h \to 0$, for $\alpha > 1$ this implies $F'(x) = 0$, and for $0 < \alpha < 1$ it is a smoothness condition stronger than continuity and weaker than differentiability.

31.19. Show (see (31.31)) that (apart from addition of constants) a function can have only one representation $F_1 + F_2$ with F_1 absolutely continuous and F_2 singular.

31.20. Show that the F_s in the Lebesgue decomposition can be further split into $F_d + F_{cs}$, where F_{cs} is continuous and singular and F_d increases only in jumps in the sense that the corresponding measure is discrete. The complete decomposition is then $F = F_{ac} + F_{cs} + F_d$.

31.21. (a) Suppose that $x_1 < x_2 < \cdots$ and $\sum_n |F(x_n)| = \infty$. Show that, if F assumes the value 0 in each interval (x_n, x_{n+1}), then it is of unbounded variation.

(b) Define F over $[0, 1]$ by $F(0) = 0$ and $F(x) = x^\alpha \sin x^{-1}$ for $x > 0$. For which values of α is F of bounded variation?

31.22. 14.4↑ If f is nonnegative and Lebesgue integrable, then by Theorem 31.3 and (31.8), except for x in a set of Lebesgue measure 0,

$$(31.35) \qquad \frac{1}{v-u} \int_u^v f(t)\, dt \to f(x)$$

if $u \le x \le v$, $u < v$, and $u, v \to x$. There is an analogue in which Lebesgue measure is replaced by a general probability measure μ: If f is nonnegative and integrable with respect to μ, then as $h \downarrow 0$,

$$(31.36) \qquad \frac{1}{\mu(x-h, x+h]} \int_{(x-h, x+h]} f(t)\mu(dt) \to f(x)$$

on a set of μ-measure 1. Let F be the distribution function corresponding to μ, and put $\varphi(u) = \inf[x: u \le F(x)]$ for $0 < u < 1$ (see (14.5)). Deduce (31.36) from (31.35) by change of variable and Problem 14.4.

SECTION 32. THE RADON-NIKODYM THEOREM

If f is a nonnegative function on a measure space $(\Omega, \mathscr{F}, \mu)$, then $\nu(A) = \int_A f\, d\mu$ defines another measure on $\mathscr{F}$. In the terminology of Section 16, ν has *density* f with respect to μ; see (16.11). For each A in $\mathscr{F}$, $\mu(A) = 0$ implies that $\nu(A) = 0$. The purpose of this section is to show conversely that if this last condition holds and ν and μ are σ-finite on $\mathscr{F}$, then ν has a density with respect to μ. This was proved for the case $(R^1, \mathscr{R}^1, \lambda)$ in Theorems 31.7 and 31.8. The theory of the preceding section, although illuminating, is not required here.

Additive Set Functions

Throughout this section, $(\Omega, \mathscr{F})$ is a measurable space. All sets involved are assumed as usual to lie in $\mathscr{F}$.

An *additive set function* is a function φ from $\mathscr{F}$ to the reals for which

$$(32.1) \qquad \varphi\left(\bigcup_n A_n \right) = \sum_n \varphi(A_n)$$

if $A_1, A_2, \ldots$ is a finite or infinite sequence of disjoint sets. A set function differs from a measure in that the values $\varphi(A)$ may be negative but must be finite—the special values $+\infty$ and $-\infty$ are prohibited. It will turn out that the series on the right in (32.1) must in fact converge absolutely, but this need not be assumed. Note that $\varphi(\varnothing) = 0$.

Example 32.1. If μ_1 and μ_2 are finite measures, then $\varphi(A) = \mu_1(A) - \mu_2(A)$ is an additive set function. It will turn out that the general additive set function has this form. A special case of this if $\varphi(A) = \int_A f \, d\mu$, where f is integrable (not necessarily nonnegative). ∎

The proof of the main theorem of this section (Theorem 32.2) requires certain facts about additive set functions, even though the statement of the theorem involves only measures.

Lemma 1. *If $E_u \uparrow E$ or $E_u \downarrow E$, then $\varphi(E_u) \to \varphi(E)$.*

Proof. If $E_u \uparrow E$, then $\varphi(E) = \varphi(E_1 \cup \bigcup_{u=1}^{\infty}(E_{u+1} - E_u)) = \varphi(E_1) + \sum_{u=1}^{\infty}\varphi(E_{u+1} - E_u) = \lim_u[\varphi(E_1) + \sum_{u=1}^{u-1}\varphi(E_{u+1} - E_u)] = \lim_u \varphi(E_u)$ by (32.1). If $E_u \downarrow E$, then $E_u^c \uparrow E^c$, and hence $\varphi(E_u) = \varphi(\Omega) - \varphi(E_u^c) \to \varphi(\Omega) - \varphi(E^c) = \varphi(E)$. ∎

Although this result is essentially the same as the corresponding ones for measures, it does require separate proof. Note that the limits need not be monotone unless φ happens to be a measure.

The Hahn Decomposition

Theorem 32.1. *For any additive set function φ, there exist disjoint sets A^+ and A^- such that $A^+ \cup A^- = \Omega$, $\varphi(E) \geq 0$ for all E in A^+, and $\varphi(E) \leq 0$ for all E in A^-.*

A set A is *positive* if $\varphi(E) \geq 0$ for $E \subset A$ and *negative* if $\varphi(E) \leq 0$ for $E \subset A$. The A^+ and A^- in the theorem decompose Ω into a positive and a negative set. This is the *Hahn decomposition*.

If $\varphi(A) = \int_A f \, d\mu$ (see Example 32.1), the result is easy: take $A^+ = [f \geq 0]$ and $A^- = [f < 0]$.

PROOF. Let $\alpha = \sup[\varphi(A): A \in \mathscr{F}]$. Suppose that there exists a set A^+ satisfying $\varphi(A^+) = \alpha$ (which implies that α is finite). Let $A^- = \Omega - A^+$. If $A \subset A^+$ and $\varphi(A) < 0$, then $\varphi(A^+ - A) > \alpha$, an impossibility; hence A^+ is a positive set. If $A \subset A^-$ and $\varphi(A) > 0$, then $\varphi(A^+ \cup A) > \alpha$, an impossibility; hence A^- is a negative set.

It is therefore only necessary to construct a set A^+ for which $\varphi(A^+) = \alpha$. Choose sets A_n such that $\varphi(A_n) \to \alpha$, and let $A = \bigcup_n A_n$. For each n consider the 2^n sets B_{ni} (some perhaps empty) that are intersections of the form $\bigcap_{k=1}^{n} A'_k$, where each A'_k is either A_k or else $A - A_k$. The collection $\mathscr{B}_n = [B_{ni}: 1 \le i \le 2^n]$ of these sets partitions A. Clearly, $\mathscr{B}_n$ refines $\mathscr{B}_{n-1}$: each B_{nj} is contained in exactly one of the $B_{n-1,i}$.

Let C_n be the union of those B_{ni} in $\mathscr{B}_n$ for which $\varphi(B_{ni}) > 0$. Since A_n is the union of certain of the B_{ni}, it follows that $\varphi(A_n) \le \varphi(C_n)$. Since the partitions $\mathscr{B}_1, \mathscr{B}_2, \ldots$ are successively finer, $m < n$ implies that $(C_m \cup \cdots \cup C_{n-1} \cup C_n) - (C_m \cup \cdots \cup C_{n-1})$ is the union (perhaps empty) of certain of the sets B_{ni}; the B_{ni} in this union must satisfy $\varphi(B_{ni}) > 0$ because they are contained in C_n. Therefore, $\varphi(C_m \cup \cdots \cup C_{n-1}) \le \varphi(C_m \cup \cdots \cup C_n)$, so that by induction $\varphi(A_m) \le \varphi(C_m) \le \varphi(C_m \cup \cdots \cup C_n)$. If $D_m = \bigcup_{n=m}^{\infty} C_n$, then by Lemma 1 (take $E_t = C_m \cup \cdots \cup C_{m+t}$) $\varphi(A_m) \le \varphi(D_m)$. Let $A^+ = \bigcap_{m=1}^{\infty} D_m$ (note that $A^+ = \limsup_n C_n$), so that $D_m \downarrow A^+$. By Lemma 1, $\alpha = \lim_m \varphi(A_m) \le \lim_m \varphi(D_m) = \varphi(A^+)$. Thus A^+ does have maximal φ-value. ∎

If $\varphi^+(A) = \varphi(A \cap A^+)$ and $\varphi^-(A) = -\varphi(A \cap A^-)$, then φ^+ and φ^- are finite measures. Thus

$$(32.2) \qquad \varphi(A) = \varphi^+(A) - \varphi^-(A)$$

represents the set function φ as the difference of two finite measures having disjoint supports. If $E \subset A$, then $\varphi(E) \le \varphi^+(E) \le \varphi^+(A)$, and there is equality if $E = A \cap A^+$. Therefore, $\varphi^+(A) = \sup_{E \subset A} \varphi(E)$. Similarly, $\varphi^-(A) = -\inf_{E \subset A} \varphi(E)$. The measures φ^+ and φ^- are called the *upper* and *lower* *variations* of φ, and the measure $|\varphi|$ with value $\varphi^+(A) + \varphi^-(A)$ at A is called the *total variation*. The representation (32.2) is the *Jordan decomposition*.

Absolute Continuity and Singularity

Measures μ and ν on $(\Omega, \mathscr{F})$ are by definition *mutually singular* if they have disjoint supports—that is, if there exist sets S_μ and S_ν such that

$$(32.3) \qquad \begin{cases} \mu(\Omega - S_\mu) = 0, \quad \nu(\Omega - S_\nu) = 0, \\ S_\mu \cap S_\nu = \varnothing. \end{cases}$$

In this case μ is also said to be *singular with respect to* ν and ν singular with respect to μ. Note that measures are automatically singular if one of them is identically 0.

According to Theorem 31.5 a finite measure on R^1 with distribution function F is singular with respect to Lebesgue measure in the sense of (32.3) if and only if $F'(x) = 0$ except on a set of Lebesgue measure 0. In Section 31 the latter condition was taken as the definition of singularity, but of course it is the requirement of disjoint supports that can be generalized from R^1 to an arbitrary Ω.

The measure ν is *absolutely continuous* with respect to μ if for each A in $\mathscr{F}$, $\mu(A) = 0$ implies $\nu(A) = 0$. In this case ν is also said to be *dominated* by μ, and the relation is indicated by $\nu \ll \mu$. If $\nu \ll \mu$ and $\mu \ll \nu$, the measures are *equivalent*, indicated by $\nu \equiv \mu$.

A finite measure on the line is by Theorem 31.7 absolutely continuous in this sense with respect to Lebesgue measure if and only if the corresponding distribution function F satisfies the condition (31.28). The latter condition, taken in Section 31 as the definition of absolute continuity, is again not the one that generalizes from R^1 to Ω.

There is an ϵ–δ idea related to the definition of absolute continuity given above. Suppose that for every ϵ there exists a δ such that

$$(32.4) \qquad\qquad \nu(A) < \epsilon \qquad \text{if } \mu(A) < \delta.$$

If this condition holds, $\mu(A) = 0$ implies that $\nu(A) < \epsilon$ for all ϵ, and so $\nu \ll \mu$. Suppose, on the other hand, that this condition fails and that ν is finite. Then for some ϵ there exist sets A_n such that $\mu(A_n) < n^{-2}$ and $\nu(A_n) \geq \epsilon$. If $A = \limsup_n A_n$, then $\mu(A) = 0$ by the first Borel–Cantelli lemma (which applies to arbitrary measures), but $\nu(A) \geq \epsilon > 0$ by the right-hand inequality in (4.9) (which applies because ν is finite). Hence $\nu \ll \mu$ fails, and so (32.4) follows if ν is finite and $\nu \ll \mu$. *If ν is finite, in order that* $\nu \ll \mu$ *it is therefore necessary and sufficient that for every* ϵ *there exist a* δ *satisfying* (32.4). This condition is not suitable as a definition, because it need not follow from $\nu \ll \mu$ if ν is infinite.[†]

The Main Theorem

If $\nu(A) = \int_A f \, d\mu$, then certainly $\nu \ll \mu$. The *Radon–Nikodym theorem* goes in the opposite direction:

Theorem 32.2. *If* μ *and* ν *are* σ-*finite measures such that* $\nu \ll \mu$, *then there exists a nonnegative* f, *a density, such that* $\nu(A) = \int_A f \, d\mu$ *for all* $A \in \mathscr{F}$. *For two such densities* f *and* g, $\mu[f \neq g] = 0$.

[†]See Problem 32.3.

The uniqueness of the density up to sets of μ-measure 0 is settled by Theorem 16.10. It is only the existence that must be proved.

The density f is integrable μ if and only if ν is finite. But since f is integrable μ over A if $\nu(A) < \infty$, and since ν is assumed σ-finite, $f < \infty$ except on a set of μ-measure 0; and f can be taken finite everywhere. By Theorem 16.11, integrals with respect to ν can be calculated by the formula

$$(32.5) \qquad \int_A h \, d\nu = \int_A hf \, d\mu.$$

The density whose existence is to be proved is called the *Radon–Nikodym derivative* of ν with respect to μ and is often denoted $d\nu/d\mu$. The term *derivative* is appropriate because of Theorems 31.3 and 31.8: For an absolutely continuous distribution function F on the line, the corresponding measure μ has with respect to Lebesgue measure the Radon–Nikodym derivative F'. Note that (32.5) can be written

$$(32.6) \qquad \int_A h \, d\nu = \int_A h \frac{d\nu}{d\mu} \, d\mu.$$

Suppose that Theorem 32.2 holds for finite μ and ν (which is in fact enough for the probabilistic applications in the sections that follow). In the σ-finite case there is a countable decomposition of Ω into $\mathscr{F}$-sets A_n for which $\mu(A_n)$ and $\nu(A_n)$ are both finite. If

$$(32.7) \qquad \mu_n(A) = \mu(A \cap A_n), \qquad \nu_n(A) = \nu(A \cap A_n),$$

then $\nu \ll \mu$ implies $\nu_n \ll \mu_n$, and so $\nu_n(A) = \int_A f_n \, d\mu_n$ for some density f_n. Since μ_n has density I_{A_n} with respect to μ (Example 16.9),

$$\nu(A) = \sum_n \nu_n(A) = \sum_n \int_A f_n \, d\mu_n = \sum_n \int_A f_n I_{A_n} \, d\mu$$

$$= \int_A \sum_n f_n I_{A_n} \, d\mu.$$

Thus $\sum_n f_n I_{A_n}$ is the density sought.

It is therefore enough to treat finite μ and ν. This requires a preliminary result.

Lemma 2. *If μ and ν are finite measures and are not mutually singular, then there exists a set A and a positive ϵ such that $\mu(A) > 0$ and $\epsilon\mu(E) \le \nu(E)$ for all $E \subset A$.*

PROOF. Let $A_n^+ \cup A_n^-$ be a Hahn decomposition for the set function $\nu - n^{-1}\mu$; put $M = \bigcup_n A_n^+$, so that $M^c = \bigcap_n A_n^-$. Since M^c is in the negative set A_n^- for $\nu - n^{-1}\mu$, it follows that $\nu(M^c) \le n^{-1}\mu(M^c)$; since this holds for all n, $\nu(M^c) = 0$. Thus M supports ν, and from the fact that μ and ν are not mutually singular it follows that M^c cannot support μ—that is, that $\mu(M)$ must be positive. Therefore, $\mu(A_n^+) > 0$ for some n. Take $A = A_n^+$ and $\epsilon = n^{-1}$.   ∎

Example 32.2. Suppose that $(\Omega, \mathscr{F}) = (R^1, \mathscr{R}^1)$, μ is Lebesgue measure λ, and $\nu(a, b] = F(b) - F(a)$. If ν and λ do not have disjoint supports, then by Theorem 31.5, $\lambda[x: F'(x) > 0] > 0$ and hence for some ϵ, $A = [x: F'(x) > \epsilon]$ satisfies $\lambda(A) > 0$. If $E = (a, b]$ is a sufficiently small interval about an x in A, then $\nu(E)/\lambda(E) = (F(b) - F(a))/(b - a) \ge \epsilon$, which is the same thing as $\epsilon\lambda(E) \le \nu(E)$.   ∎

Thus Lemma 2 ties in with derivatives and quotients $\nu(E)/\mu(E)$ for "small" sets E. Martingale theory links Radon–Nikodym derivatives with such quotients; see Theorem 35.7 and Example 35.10.

PROOF OF THEOREM 32.2. Suppose that μ and ν are finite measures satisfying $\nu \ll \mu$. Let $\mathscr{G}$ be the class of nonnegative functions g such that $\int_E g\, d\mu \le \nu(E)$ for all E. If g and g' lie in $\mathscr{G}$, then $\max(g, g')$ also lies in $\mathscr{G}$ because

$$\int_E \max(g, g')\, d\mu = \int_{E \cap [g \ge g']} g\, d\mu + \int_{E \cap [g' > g]} g'\, d\mu$$

$$\le \nu(E \cap [g \ge g']) + \nu(E \cap [g' > g]) = \nu(E).$$

Thus $\mathscr{G}$ is closed under the formation of finite maxima. Suppose that functions g_n lie in $\mathscr{G}$ and $g_n \uparrow g$. Then $\int_E g\, d\mu = \lim_n \int_E g_n\, d\mu \le \nu(E)$ by the monotone convergence theorem, so that g lies in $\mathscr{G}$. Thus $\mathscr{G}$ is closed under nondecreasing passages to the limit.

Let $\alpha = \sup \int g\, d\mu$ for g ranging over $\mathscr{G}$ ($\alpha \le \nu(\Omega)$). Choose g_n in $\mathscr{G}$ so that $\int g_n\, d\mu > \alpha - n^{-1}$. If $f_n = \max(g_1, \ldots, g_n)$ and $f = \lim f_n$, then f lies in $\mathscr{G}$ and $\int f\, d\mu = \lim_n \int f_n\, d\mu \ge \lim_n \int g_n\, d\mu = \alpha$. Thus f is an element of $\mathscr{G}$ for which $\int f\, d\mu$ is maximal.

Define ν_{ac} by $\nu_{ac}(E) = \int_E f\, d\mu$ and ν_s by $\nu_s(E) = \nu(E) - \nu_{ac}(E)$. Thus

(32.8) $\nu(E) = \nu_{ac}(E) + \nu_s(E) = \int_E f\, d\mu + \nu_s(E).$

Since f is in $\mathscr{G}$, ν_s as well as ν_{ac} is a finite measure—that is, nonnegative. Of course, ν_{ac} is absolutely continuous with respect to μ.

Suppose that ν_s fails to be singular with respect to μ. It then follows from Lemma 2 that there are a set A and a positive ϵ such that $\mu(A) > 0$ and $\epsilon\mu(E) \le \nu_s(E)$ for all $E \subset A$. Then for every E

$$\int_E (f + \epsilon I_A) \, d\mu = \int_E f \, d\mu + \epsilon\mu(E \cap A) \le \int_E f \, d\mu + \nu_s(E \cap A)$$

$$= \int_{E \cap A} f \, d\mu + \nu_s(E \cap A) + \int_{E - A} f \, d\mu$$

$$= \nu(E \cap A) + \int_{E - A} f \, d\mu \le \nu(E \cap A) + \nu(E - A)$$

$$= \nu(E).$$

In other words, $f + \epsilon I_A$ lies in $\mathscr{S}$; since $\int (f + \epsilon I_A) \, d\mu = \alpha + \epsilon\mu(A) > \alpha$, this contradicts the maximality of f.

Therefore, μ and ν_s are mutually singular, and there exists an S such that $\nu_s(S) = \mu(S^c) = 0$. But since $\nu \ll \mu$, $\nu_s(S^c) \le \nu(S^c) = 0$, and so $\nu_s(\Omega) = 0$. The rightmost term in (32.8) thus drops out. ∎

Absolute continuity was not used until the last step of the proof, and what the argument shows is that ν always has a decomposition (32.8) into an *absolutely continuous part* and a *singular part* with respect to μ. This is the *Lebesgue decomposition*, and it generalizes the one in the preceding section (see (31.31)).

PROBLEMS

32.1. There are two ways to show that the convergence in (32.1) must be absolute: Use the Jordan decomposition. Use the fact that a series converges absolutely if it has the same sum no matter what order the terms are taken in.

32.2. If $A^+ \cup A^-$ is a Hahn decomposition of φ, there may be other ones $A_1^+ \cup A_1^-$. Construct an example of this. Show that there is uniqueness to the extent that $\varphi(A^+ \triangle A_1^+) = \varphi(A^- \triangle A_1^-) = 0$.

32.3. Show that absolute continuity does not imply the ϵ-δ condition (32.4) if ν is infinite. *Hint.* Let $\mathscr{F}$ consist of all subsets of the space of integers, let ν be counting measure, and let μ have mass n^{-2} at n. Note that μ is finite and ν is σ-finite.

32.4. Show that the Radon–Nikodym theorem fails if μ is not σ-finite, even if ν is finite. *Hint:* Let $\mathscr{F}$ consist of the countable and the cocountable sets in an uncountable Ω, let μ be counting measure, and let $\nu(A)$ be 0 or 1 as A is countable or cocountable.

32.5. Let μ be the restriction of planar Lebesgue measure λ_2 to the σ-field $\mathscr{F} = \{A \times R^1 : A \in \mathscr{R}^1\}$ of vertical strips. Define ν on $\mathscr{F}$ by $\nu(A \times R^1) = \lambda_2(A \times (0,1))$. Show that ν is absolutely continuous with respect to μ but has no density. Why does this not contradict the Radon–Nikodym theorem?

32.6. Let μ, $\overset{\ast}{\nu}$, and ρ be σ-finite measures on $(\Omega, \mathscr{F})$. Assume the Radon–Nikodym derivatives here are everywhere nonnegative and finite.

(a) Show that $\nu \ll \mu$ and $\mu \ll \rho$ imply that $\nu \ll \rho$ and

$$\frac{d\nu}{d\rho} = \frac{d\nu}{d\mu}\frac{d\mu}{d\rho}.$$

(b) Show that $\nu \equiv \mu$ implies

$$\frac{d\nu}{d\mu} = I_{[d\mu/d\nu > 0]}\left(\frac{d\mu}{d\nu}\right)^{-1}.$$

(c) Suppose that $\mu \ll \rho$ and $\nu \ll \rho$, and let A be the set where $d\nu/d\rho > 0 = d\mu/d\rho$. Show that $\nu \ll \mu$ if and only if $\rho(A) = 0$, in which case

$$\frac{d\nu}{d\mu} = I_{[d\mu/d\rho > 0]}\frac{d\nu/d\rho}{d\mu/d\rho}.$$

32.7. Show that there is a Lebesgue decomposition (32.8) in the σ-finite as well as the finite case. Prove that it is unique.

32.8. The Radon–Nikodym theorem holds if μ is σ-finite, even if ν is not. Assume at first that μ is finite (and $\nu \ll \mu$).

(a) Let $\mathscr{B}$ be the class of ($\mathscr{F}$-sets) B such that $\mu(E) = 0$ or $\nu(E) = \infty$ for each $E \subset B$. Show that $\mathscr{B}$ contains a set B_0 of maximal μ-measure.

(b) Let $\mathscr{C}$ be the class of sets in $\Omega_0 = B_0^c$ that are countable unions of sets of finite ν-measure. Show that $\mathscr{C}$ contains a set C_0 of maximal μ-measure. Let $D_0 = \Omega_0 - C_0$.

(c) Deduce from the maximality of B_0 and C_0 that $\mu(D_0) = \nu(D_0) = 0$.

(d) Let $\nu_0(A) = \nu(A \cap \Omega_0)$. Using the Radon–Nikodym theorem for the pair μ, ν_0, prove it for μ, ν.

(e) Now show that the theorem holds if μ is merely σ-finite.

(f) Show that if the density can be taken everywhere finite, then ν is σ-finite.

32.9. Let μ and ν be finite measures on $(\Omega, \mathscr{F})$, and suppose that $\mathscr{F}^\circ$ is a σ-field contained in $\mathscr{F}$. Then the restrictions μ° and ν° of μ and ν to $\mathscr{F}^\circ$ are measures on $(\Omega, \mathscr{F}^\circ)$. Let $\nu_{ac}, \nu_s, \nu_{ac}^\circ, \nu_s^\circ$ be, respectively, the absolutely continuous and singular parts of ν and ν° with respect to μ and μ°. Show that $\nu_{ac}^\circ(E) \geq \nu_{ac}(E)$ and $\nu_s^\circ(E) \leq \nu_s(E)$ for $E \in \mathscr{F}^\circ$.

32.10. Suppose that μ, ν, ν_n are finite measures on $(\Omega, \mathscr{F})$ and that $\nu(A) = \Sigma_n \nu_n(A)$ for all A. Let $\nu_n(A) = \int_A f_n \, d\mu + \nu_n'(A)$ and $\nu(A) = \int_A f \, d\mu + \nu'(A)$ be the decompositions (32.8); here ν' and ν_n' are singular with respect to μ. Show that $f = \Sigma_n f_n$ except on a set of μ-measure 0 and that $\nu'(A) = \Sigma_n \nu_n'(A)$ for all A. Show that $\nu \ll \mu$ if and only if $\nu_n \ll \mu$ for all n.

32.11. 32.2↑ Absolute continuity of a set function φ with respect to a measure μ is defined just as if φ were itself a measure: $\mu(A) = 0$ must imply that $\varphi(A) = 0$. Show that, if this holds and μ is σ-finite, then $\varphi(A) = \int_A f \, d\mu$ for some integrable f. Show that $A^+ = [\omega: f(\omega) \geq 0]$ and $A^- = [\omega: f(\omega) < 0]$ give a Hahn decomposition for φ. Show that the three variations satisfy $\varphi^+(A) = \int_A f^+ \, d\mu$, $\varphi^-(A) = \int_A f^- \, d\mu$, and $|\varphi|(A) = \int_A |f| \, d\mu$. *Hint:* To construct f, start with (32.2).

32.12. ↑ A *signed measure* φ is a set function that satisfies (32.1) if $A_1, A_2, \ldots$ are disjoint and may assume one of the values $+\infty$ and $-\infty$ but not both. Extend the Hahn and Jordan decompositions to signed measures

32.13. 31.22↑ Suppose that μ and ν are a probability measure and a σ-finite measure on the line and that $\nu \ll \mu$. Show that the Radon–Nikodym derivative f satisfies

$$\lim_{h \to 0} \frac{\nu(x-h, x+h]}{\mu(x-h, x+h]} = f(x)$$

on a set of μ-measure 1.

32.14. Find on the unit interval uncountably many probability measures μ_p, $0 < p < 1$, with supports S_p such that $\mu_p\{x\} = 0$ for each x and p and the S_p are disjoint in pairs.

32.15. Let $\mathscr{F}_0$ be the field consisting of the finite and the cofinite sets in an uncountable Ω. Define φ on $\mathscr{F}_0$ by taking $\varphi(A)$ to be the number of points in A if A is finite, and the negative of the number of points in A^c if A is cofinite. Show that (32.1) holds (this is not true if Ω is countable). Show that there are no negative sets for φ (except the empty set), that there is no Hahn decomposition, and that φ does not have bounded range.

SECTION 33. CONDITIONAL PROBABILITY

The concepts of conditional probability and expected value with respect to a σ-field underlie much of modern probability theory. The difficulty in understanding these ideas has to do not with mathematical detail so much as with probabilistic meaning, and the way to get at this meaning is through calculations and examples, of which there are many in this section and the next.

The Discrete Case

Consider first the conditional probability of a set A with respect to another set B. It is defined of course by $P(A|B) = P(A \cap B)/P(B)$, unless $P(B)$ vanishes, in which case it is not defined at all.

It is helpful to consider conditional probability in terms of an observer in possession of partial information.[†] A probability space $(\Omega, \mathscr{F}, P)$ describes

---

[†] As always, *observer*, *information*, *know*, and so on are informal, nonmathematical terms; see the related discussion in Section 4 (p. 57).

the working of a mechanism, governed by chance, which produces a result ω distributed according to P; $P(A)$ is for the observer the probability that the point ω produced lies in A. Suppose now that ω lies in B and that the observer learns this fact and no more. From the point of view of the observer, now in possession of this partial information about ω, the probability that ω also lies in A is $P(A|B)$ rather than $P(A)$. This is the idea lying back of the definition.

If, on the other hand, ω happens to lie in B^c and the observer learns of this, his probability instead becomes $P(A|B^c)$. These two conditional probabilities can be linked together by the simple function

$$(33.1) \qquad f(\omega) = \begin{cases} P(A|B) & \text{if } \omega \in B, \\ P(A|B^c) & \text{if } \omega \in B^c. \end{cases}$$

The observer learns whether ω lies in B or in B^c; his new probability for the event $\omega \in A$ is then just $f(\omega)$. Although the observer does not in general know the argument ω of f, he can calculate the value $f(\omega)$ because he knows which of B and B^c contains ω. (Note conversely that from the value $f(\omega)$ it is possible to determine whether ω lies in B or in B^c, unless $P(A|B) = P(A|B^c)$—that is, unless A and B are independent, in which case the conditional probability coincides with the unconditional one anyway.)

The sets B and B^c partition Ω, and these ideas carry over to the general partition. Let $B_1, B_2, \ldots$ be a finite or countable partition of Ω into $\mathscr{F}$-sets, and let $\mathscr{G}$ consist of all the unions of the B_i. Then $\mathscr{G}$ is the σ-field generated by the B_i. For A in $\mathscr{F}$, consider the function with values

$$(33.2) \quad f(\omega) = P(A|B_i) = \frac{P(A \cap B_i)}{P(B_i)} \quad \text{if } \omega \in B_i, \qquad i = 1, 2, \ldots.$$

If the observer learns which element B_i of the partition it is that contains ω, then his new probability for the event $\omega \in A$ is $f(\omega)$. The partition $\{B_i\}$, or equivalently the σ-field $\mathscr{G}$, can be regarded as an experiment, and to learn which B_i it is that contains ω is to learn the outcome of the experiment. For this reason the function or random variable f defined by (33.2) is called the *conditional probability of A given $\mathscr{G}$* and is denoted $P[A\|\mathscr{G}]$. This is written $P[A\|\mathscr{G}]_\omega$ whenever the argument ω needs to be explicitly shown.

Thus $P[A\|\mathscr{G}]$ is the function whose value on B_i is the ordinary conditional probability $P(A|B_i)$. This definition needs to be completed, because $P(A|B_i)$ is not defined if $P(B_i) = 0$. In this case $P[A\|\mathscr{G}]$ will be taken to have any constant value on B_i; the value is arbitrary but must be the same over all of the set B_i. If there are nonempty sets B_i for which $P(B_i) = 0$, $P[A\|\mathscr{G}]$ therefore stands for any one of a family of functions on Ω. A specific such function is for emphasis often called a *version* of the conditional

probability. Note that any two versions are equal except on a set of probability 0.

Example 33.1. Consider the Poisson process. Suppose that $0 \le s \le t$, and let $A = [N_s = 0]$ and $B_i = [N_t = i]$, $i = 0, 1, \ldots$. Since the increments are independent (Section 23), $P(A|B_i) = P[N_s = 0]P[N_t - N_s = i]/P[N_t = i]$, and since they have Poisson distributions (see (23.9)), a simple calculation reduces this to

$$(33.3) \qquad P[N_s = 0 \| \mathscr{I}]_\omega = \left(1 - \frac{s}{t}\right)^i \quad \text{if } \omega \in B_i, \qquad i = 0, 1, 2, \ldots .$$

Since $i = N_t(\omega)$ on B_i, this can be written

$$(33.4) \qquad P[N_s = 0 \| \mathscr{I}]_\omega = \left(1 - \frac{s}{t}\right)^{N_t(\omega)}.$$

Here the experiment or observation corresponding to $\{B_i\}$ or $\mathscr{I}$ determines the number of events—telephone calls, say—occurring in the time interval $[0, t]$. For an observer who knows this number but not the locations of the calls within $[0, t]$, (33.4) gives his probability for the event that none of them occurred before time s. Although this observer does not known ω, he knows $N_t(\omega)$, which is all he needs to calculate the right side of (33.4). ∎

Example 33.2. Suppose that $X_0, X_1, \ldots$ is a Markov chain with state space S as in Section 8. The events

$$(33.5) \qquad [X_0 = i_0, \ldots, X_n = i_n]$$

form a finite or countable partition of Ω as $i_0, \ldots, i_n$ range over S. If $\mathscr{I}_n$ is the σ-field generated by this partition, then by the defining condition (8.2) for Markov chains, $P[X_{n+1} = j \| \mathscr{I}_n]_\omega = p_{i_n j}$ holds for ω in (33.5). The sets

$$(33.6) \qquad [X_n = i]$$

for $i \in S$ also partition Ω, and they generate a σ-field $\mathscr{I}_n^0$ smaller than $\mathscr{I}_n$. Now (8.2) also stipulates $P[X_{n+1} = j \| \mathscr{I}_n^0]_\omega = p_{ij}$ for ω in (33.6), and the essence of the Markov property is that

$$(33.7) \qquad P[X_{n+1} = j \| \mathscr{I}_n] = P[X_{n+1} = j \| \mathscr{I}_n^0].$$ ∎

The General Case

If $\mathscr{I}$ is the σ-field generated by a partition $B_1, B_2, \ldots$, then the general element of $\mathscr{I}$ is a disjoint union $B_{i_1} \cup B_{i_2} \cup \cdots$, finite or countable, of certain of the B_i. To know which set B_i it is that contains ω is the same thing

as to know which sets in $\mathscr{G}$ contain ω and which do not. This second way of looking at the matter carries over to the general σ-field $\mathscr{G}$ contained in $\mathscr{F}$. (As always, the probability space is $(\Omega, \mathscr{F}, P)$.) The σ-field $\mathscr{G}$ will not in general come from a partition as above.

One can imagine an observer who knows for each G in $\mathscr{G}$ whether $\omega \in G$ or $\omega \in G^c$. Thus the σ-field $\mathscr{G}$ can in principle be identified with an experiment or observation. This is the point of view adopted in Section 4; see p. 57. It is natural to try and define conditional probabilities $P[A\|\mathscr{G}]$ with respect to the experiment $\mathscr{G}$. To do this, fix an A in $\mathscr{F}$ and define a finite measure ν on $\mathscr{G}$ by

$$\nu(G) = P(A \cap G), \qquad G \in \mathscr{G}.$$

Then $P(G) = 0$ implies that $\nu(G) = 0$. The Radon–Nikodym theorem can be applied to the measures ν and P on the measurable space $(\Omega, \mathscr{G})$ because the first one is absolutely continuous with respect to the second.[†] It follows that there exists a function or random variable f, measurable $\mathscr{G}$ and integrable with respect to P, such that[†] $P(A \cap G) = \nu(G) = \int_G f\,dP$ for all G in $\mathscr{G}$.

Denote this function f by $P[A\|\mathscr{G}]$. It is a random variable with two properties:

(i) $P[A\|\mathscr{G}]$ *is measurable $\mathscr{G}$ and integrable.*
(ii) $P[A\|\mathscr{G}]$ *satisfies the functional equation*

$$(33.8) \qquad \int_G P[A\|\mathscr{G}]\,dP = P(A \cap G), \qquad G \in \mathscr{G}.$$

There will in general be many such random variables $P[A\|\mathscr{G}]$, but any two of them are equal with probability 1. A specific such random variable is called a *version* of the conditional probability.

If $\mathscr{G}$ is generated by a partition $B_1, B_2, \ldots$ the function f defined by (33.2) is measurable $\mathscr{G}$ because $[\omega : f(\omega) \in H]$ is the union of those B_i over which the constant value of f lies in H. Any G in $\mathscr{G}$ is a disjoint union $G = \bigcup_k B_{i_k}$, and $P(A \cap G) = \sum_k P(A|B_{i_k})P(B_{i_k})$, so that (33.2) satisfies (33.8) as well. Thus the general definition is an extension of the one for the discrete case.

Condition (i) in the definition above in effect requires that the values of $P[A\|\mathscr{G}]$ depend only on the sets in $\mathscr{G}$. An observer who knows the outcome of $\mathscr{G}$ viewed as an experiment knows for each G in $\mathscr{G}$ whether it contains ω or not; for each x he knows this in particular for the set $[\omega' : P[A\|\mathscr{G}]_{\omega'} = x]$,

---

[†]Let P_0 be the restriction of P to $\mathscr{G}$ (Example 10.4), and find on $(\Omega, \mathscr{G})$ a density f for ν with respect to P_0. Then, for $G \in \mathscr{G}$, $\nu(G) = \int_G f\,dP_0 = \int_G f\,dP$ (Example 16.4). If g is another such density, then $P[f \neq g] = P_0[f \neq g] = 0$.

and hence he knows in principle the functional value $P[A\|\mathcal{G}]_\omega$ even if he does not know ω itself. In Example 33.1 a knowledge of $N_t(\omega)$ suffices to determine the value of (33.4)—ω itself is not needed.

Condition (ii) in the definition has a gambling interpretation. Suppose that the observer, after he has learned the outcome of $\mathcal{G}$, is offered the opportunity to bet on the event A (unless A lies in $\mathcal{G}$, he does not yet know whether or not it occurred). He is required to pay an entry fee of $P[A\|\mathcal{G}]$ units and will win 1 unit if A occurs and nothing otherwise. If the observer decides to bet and pays his fee, he gains $1 - P[A\|\mathcal{G}]$ if A occurs and $-P[A\|\mathcal{G}]$ otherwise, so that his gain is

$$(1 - P[A\|\mathcal{G}])I_A + (-P[A\|\mathcal{G}])I_{A^c} = I_A - P[A\|\mathcal{G}].$$

If he declines to bet, his gain is of course 0. Suppose that he adopts the strategy of betting if G occurs but not otherwise, where G is some set in $\mathcal{G}$. He can actually carry out this strategy, since after learning the outcome of the experiment $\mathcal{G}$ he knows whether or not G occurred. His expected gain with this strategy is his gain integrated over G:

$$\int_G (I_A - P[A\|\mathcal{G}])\, dP.$$

But (33.8) is exactly the requirement that this vanish for each G in $\mathcal{G}$. Condition (ii) requires then that each strategy be fair in the sense that the observer stands neither to win nor to lose on the average. Thus $P[A\|\mathcal{G}]$ is the just entry fee, as intuition requires.

Example 33.3. Suppose that $A \in \mathcal{G}$, which will always hold if $\mathcal{G}$ coincides with the whole σ-field $\mathcal{F}$. Then I_A satisfies conditions (i) and (ii), so that $P[A\|\mathcal{G}] = I_A$ with probability 1. If $A \in \mathcal{G}$, then to know the outcome of $\mathcal{G}$ viewed as an experiment is in particular to know whether or not A has occurred.   ∎

Example 33.4. If $\mathcal{G}$ is $\{0, \Omega\}$, the smallest possible σ-field, every function measurable $\mathcal{G}$ must be constant. Therefore, $P[A\|\mathcal{G}]_\omega = P(A)$ for all ω in this case. The observer learns nothing from the experiment $\mathcal{G}$.   ∎

According to these two examples, $P[A\|\{0, \Omega\}]$ is identically $P(A)$, whereas I_A is a version of $P[A\|\mathcal{F}]$. For any $\mathcal{G}$, the function identically equal to $P(A)$ satisfies condition (i) in the definition of conditional probability, whereas I_A satisfies condition (ii). Condition (i) becomes more stringent as $\mathcal{G}$ decreases, and condition (ii) becomes more stringent as $\mathcal{G}$ increases. The two conditions work in opposite directions and between them delimit the class of versions of $P[A\|\mathcal{G}]$.

Example 33.5. Let Ω be the plane R^2 and let $\mathscr{F}$ be the class $\mathscr{R}^2$ of planar Borel sets. A point of Ω is a pair (x, y) of reals. Let $\mathscr{G}$ be the σ-field consisting of the vertical strips, the product sets $E \times R^1 = [(x, y): x \in E]$, where E is a linear Borel set. If the observer knows for each strip $E \times R^1$ whether or not it contains (x, y), then, as he knows this for each one-point set E, he knows the value of x. Thus the experiment $\mathscr{G}$ consists in the determination of the first coordinate of the sample point. Suppose now that P is a probability measure on $\mathscr{R}^2$ having a density $f(x, y)$ with respect to planar Lebesgue measure: $P(A) = \iint_A f(x, y) \, dx \, dy$. Let A be a horizontal strip $R^1 \times F = [(x, y): y \in F]$, F being a linear Borel set. The conditional probability $P[A\|\mathscr{G}]$ can be calculated explicitly.

Put

$$(33.9) \qquad \varphi(x, y) = \frac{\displaystyle\int_F f(x, t) \, dt}{\displaystyle\int_{R^1} f(x, t) \, dt}.$$

Set $\varphi(x, y) = 0$, say, at points where the denominator here vanishes; these points form a set of P-measure 0. Since $\varphi(x, y)$ is a function of x alone, it is measurable $\mathscr{G}$. The general element of $\mathscr{G}$ being $E \times R^1$, it will follow that φ is a version of $P[A\|\mathscr{G}]$ if it is shown that

$$(33.10) \qquad \int_{E \times R^1} \varphi(x, y) \, dP(x, y) = P(A \cap (E \times R^1)).$$

Since $A = R^1 \times F$, the right side here is $P(E \times F)$. Since P has density f, Theorem 16.11 and Fubini's theorem reduce the left side to

$$\int_E \left\{ \int_{R^1} \varphi(x, y) f(x, y) \, dy \right\} dx = \int_E \left\{ \int_F f(x, t) \, dt \right\} dx$$

$$= \iint_{E \times F} f(x, y) \, dx \, dy = P(E \times F).$$

Thus (33.9) does give a version of $P[R^1 \times F\|\mathscr{G}]$. ∎

The right side of (33.9) is the classical formula for the conditional probability of the event $R^1 \times F$ (the event that $y \in F$) given the event $\{x\} \times R^1$ (given the value of x). Since the event $\{x\} \times R^1$ has probability 0, the formula $P(A|B) = P(A \cap B)/P(B)$ does not work here. The whole point of this section is the systematic development of a notion of conditional probability that covers conditioning with respect to events of probability 0. This is accomplished by conditioning with respect to *collections* of events—that is, with respect to σ-fields $\mathscr{G}$.

Example 33.6. The set A is by definition independent of the σ-field $\mathscr{G}$ if it is independent of each G in $\mathscr{G}$: $P(A \cap G) = P(A)P(G)$. This being the same thing as $P(A \cap G) = \int_G P(A)\,dP$, A is independent of $\mathscr{G}$ if and only if $P[A \| \mathscr{G}] = P(A)$ with probability 1. ∎

The σ-field $\sigma(X)$ generated by a random variable X consists of the sets $[\omega: X(\omega) \in H]$ for $H \in \mathscr{R}^1$; see Theorem 20.1. The conditional probability of A given X is defined as $P[A \| \sigma(X)]$ and is denoted $P[A \| X]$. Thus $P[A \| X] = P[A \| \sigma(X)]$ by definition. From the experiment corresponding to the σ-field $\sigma(X)$, one learns which of the sets $[\omega': X(\omega') = x]$ contains ω and hence learns the value $X(\omega)$. Example 33.5 is a case of this: take $X(x, y) = x$ for (x, y) in the sample space $\Omega = R^2$ there.

This definition applies without change to random vector, or, equivalently, to a finite set of random variables. It can be adapted to arbitrary sets of random variables as well. For any such set $[X_t, t \in T]$, the σ-field $\sigma[X_t, t \in T]$ it generates is the smallest σ-field with respect to which each X_t is measurable. It is generated by the collection of sets of the form $[\omega: X_t(\omega) \in H]$ for t in T and H in $\mathscr{R}^1$. The *conditional probability $P[A \| X_t, t \in T]$ of A with respect to this set of random variables* is by definition the conditional probability $P[A \| \sigma[X_t, t \in T]]$ of A with respect to the σ-field $\sigma[X_t, t \in T]$.

In this notation the property (33.7) of Markov chains becomes

$$(33.11) \qquad P\big[X_{n+1} = j \| X_0, \dots, X_n\big] = P\big[X_{n+1} = j \| X_n\big].$$

The conditional probability of $[X_{n+1} = j]$ is the same for someone who knows the present state X_n as for someone who knows the present state X_n and the past states $X_0, \dots, X_{n-1}$ as well.

Example 33.7. Let X and Y be random vectors of dimensions j and k, let μ be the distribution of X over R^j, and suppose that X and Y are *independent*. According to (20.30),

$$P[X \in H, (X, Y) \in J] = \int_H P[(x, Y) \in J]\mu(dx)$$

for $H \in \mathscr{R}^j$ and $J \in \mathscr{R}^{j+k}$. This is a consequence of Fubini's theorem; it has a conditional-probability interpretation. For each x in R^j put

$$(33.12) \qquad f(x) = P[(x, Y) \in J] = P[\omega': (x, Y(\omega')) \in J].$$

By Theorem 20.1(ii), $f(X(\omega))$ is measurable $\sigma(X)$, and since μ is the distribution of X, a change of variable gives

$$\int_{[X \in H]} f(X(\omega))P(d\omega) = \int_H f(x)\mu(dx) = P([(X, Y) \in J] \cap [X \in H]).$$

Since $[X \in H]$ is the general element of $\sigma(X)$, this proves that

$$(33.13) \qquad f(X(\omega)) = P[(X,Y) \in J \| X]_\omega$$

with probability 1. ∎

The fact just proved can be written

$$P[(X,Y) \in J \| X]_\omega = P[(X(\omega),Y) \in J]$$
$$= P[\omega': (X(\omega),Y(\omega')) \in J].$$

Replacing ω' by ω on the right here causes a notational collision like the one replacing y by x causes in $\int_a^b f(x,y)\,dy$.

Suppose that X and Y are independent random variables and that Y has distribution function F. For $J = [(u,v): \max\{u,v\} \le m]$, (33.12) is 0 for $m < x$ and $F(m)$ for $m \ge x$; if $M = \max\{X,Y\}$, then (33.13) gives

$$(33.14) \qquad P[M \le m \| X]_\omega = I_{[X \le m]}(\omega) F(m)$$

with probability 1. All equations involving conditional probabilities must be qualified in this way by the phrase *with probability 1*, because the conditional probability is unique only to within a set of probability 0.

The following theorem is useful for checking conditional probabilities.

Theorem 33.1. *Let $\mathscr{P}$ be a π-system generating the σ-field $\mathscr{G}$, and suppose that Ω is a finite or countable union of sets in $\mathscr{P}$. An integrable function f is a version of $P[A \| \mathscr{G}]$ if it is measurable $\mathscr{G}$ and if*

$$(33.15) \qquad \int_G f\,dP = P(A \cap G)$$

holds for all G in $\mathscr{P}$.

PROOF. Apply Theorem 10.4. ∎

The condition that Ω is a finite or countable union of $\mathscr{P}$-sets cannot be suppressed; see Example 10.5.

Example 33.8. Suppose that X and Y are independent random variables with a common distribution function F that is positive and continuous. What is the conditional probability of $[X \le x]$ given the random variable $M = \max\{X,Y\}$? As it should clearly be 1 if $M \le x$, suppose that $M > x$. Since $X \le x$ requires $M = Y$, the chance of which is $\frac{1}{2}$ by symmetry, the conditional probability of $[X \le x]$ should by independence be $\frac{1}{2}F(x)/F(m) = \frac{1}{2}P[X \le x | X \le m]$ with the random variable M substituted

for m. Intuition thus gives

$$(33.16) \qquad P[X \leq x \| M]_\omega = I_{[M \leq x]}(\omega) + \frac{1}{2} I_{[M > x]}(\omega) \frac{F(x)}{F(M(\omega))}.$$

It suffices to check (33.15) for sets $G = [M \leq m]$, because these form a π-system generating $\sigma(M)$. The functional equation reduces to

$$(33.17) \qquad P[M \leq \min\{x, m\}] + \frac{1}{2} \int_{x < M \leq m} \frac{F(x)}{F(M)} dP = P[M \leq m, X \leq x].$$

Since the other case is easy, suppose that $x < m$. Since the distribution of (X, Y) is product measure, it follows by Fubini's theorem and the assumed continuity of F that

$$\int_{x < M \leq m} \frac{1}{F(M)} dP = \iint_{\substack{u \leq l \\ x < l \leq m}} \frac{1}{F(v)} dF(u) dF(v)$$

$$+ \iint_{\substack{l < u \\ x < u \leq m}} \frac{1}{F(u)} dF(u) \, dF(v) = 2(F(m) - F(x)),$$

which gives (33.17). ∎

Example 33.9. A collection $[X_t : t \geq 0]$ of random variables is a *Markov process in continuous time* if for $k \geq 1$, $0 \leq t_1 \leq \cdots \leq t_k \leq u$, and $H \in \mathscr{R}^1$,

$$(33.18) \qquad P[X_u \in H \| X_{t_1}, \ldots, X_{t_k}] = P[X_u \in H \| X_{t_k}]$$

holds with probability 1. The analogue for discrete time is (33.11). (The X_n there have countable range as well, and the transition probabilities are constant in time, conditions that are not imposed here.)

Suppose that $t \leq u$. Looking on the right side of (33.18) as a version of the conditional probability on the left shows that

$$(33.19) \qquad \int_G P[X_u \in H \| X_t] \, dP = P([X_u \in H] \cap G)$$

if $0 \leq t_1 \leq \cdots \leq t_k = t \leq u$ and $G \in \sigma(X_{t_1}, \ldots, X_{t_k})$. Fix t, u, and H, and let k and $t_1, \ldots, t_k$ vary. Consider the class $\mathscr{P} = \bigcup \sigma(X_{t_1}, \ldots, X_{t_k})$, the union extending over all $k \geq 1$ and all k-tuples satisfying $0 \leq t_1 \leq \cdots \leq t_k = t$. If $A \in \sigma(X_{t_1}, \ldots, X_{t_k})$ and $B \in \sigma(X_{s_1}, \ldots, X_{s_j})$, then $A \cap B \in \sigma(X_{r_1}, \ldots, X_{r_j})$, where the r_α are the s_β and the t_γ merged together. Thus $\mathscr{P}$ is a π-system. Since $\mathscr{P}$ generates $\sigma[X_s : s \leq t]$ and $P[X_u \in H \| X_t]$ is measurable with respect to this σ-field, it follows by (33.19) and Theorem 33.1 that $P[X_u \in H \| X_t]$ is a version of $P[X_u \in H \| X_s, s \leq t]$:

$$(33.20) \qquad P[X_u \in H \| X_s, s \leq t] = P[X_u \in H \| X_t], \qquad t \leq u,$$

with probability 1.

This says that for calculating conditional probabilities about the future, the present $\sigma(X_t)$ is equivalent to the present and the *entire* past $\sigma[X_s: s \leq t]$. This follows from the apparently weaker condition (33.18). ∎

Example 33.10. The Poisson process $[N_t: t \geq 0]$ has independent increments (Section 23). Suppose that $0 \leq t_1 \leq \cdots \leq t_k \leq u$. The random vector $(N_{t_1}, N_{t_2} - N_{t_1}, \ldots, N_{t_k} - N_{t_{k-1}})$ is independent of $N_u - N_{t_k}$, and so (Theorem 20.2) $(N_{t_1}, N_{t_2}, \ldots, N_{t_k})$ is independent of $N_u - N_{t_k}$. If J is the set of points $(x_1, \ldots, x_k, y)$ in R^{k+1} such that $x_k + y \in H$, where $H \in \mathscr{R}^1$, and if ν is the distribution of $N_u - N_{t_k}$, then (33.12) is $P[(x_1, \ldots, x_k, N_u - N_{t_k}) \in J] = P[x_k + N_u - N_{t_k} \in H] = \nu(H - x_k)$. Therefore, (33.13) gives $P[N_u \in H \| N_{t_1}, \ldots, N_{t_k}] = \nu(H - N_{t_k})$. This holds also if $k = 1$, and hence $P[N_u \in H \| N_{t_1}, \ldots, N_{t_k}] = P[N_u \in H \| N_{t_k}]$. The Poisson process thus has the Markov property (33.18); this is a consequence solely of the independence of the increments. The extended Markov property (33.20) follows. ∎

Properties of Conditional Probability

Theorem 33.2. *With probability* 1, $P[\varnothing \| \mathscr{G}] = 0$, $P[\Omega \| \mathscr{G}] = 1$; *and*

$$(33.21) \qquad\qquad 0 \leq P[A \| \mathscr{G}] \leq 1$$

for each A. If $A_1, A_2, \ldots$ is a finite or countable sequence of disjoint sets, then

$$(33.22) \qquad\qquad P\left[\bigcup_n A_n \| \mathscr{G} \right] = \sum_n P[A_n \| \mathscr{G}]$$

with probability 1.

PROOF. For each version of the conditional probability, $\int_G P[A \| \mathscr{G}] \, dP = P(A \cap G) \geq 0$ for each G in $\mathscr{G}$; since $P[A \| \mathscr{G}]$ is measurable $\mathscr{G}$, it must be nonnegative except on a set of P-measure 0. The other inequality in (33.21) is proved the same way.

If the A_n are disjoint and if G lies in $\mathscr{G}$, it follows (Theorem 16.6) that

$$\int_G \left(\sum_n P[A_n \| \mathscr{G}] \right) dP = \sum_n \int_G P[A_n \| \mathscr{G}] \, dP = \sum_n P(A_n \cap G)$$

$$= P\left(\left(\bigcup_n A_n \right) \cap G \right).$$

Thus $\sum_n P[A_n \| \mathscr{G}]$, which is certainly measurable $\mathscr{G}$, satisfies the functional equation for $P[\bigcup_n A_n \| \mathscr{G}]$, and so must coincide with it except perhaps on a set of P-measure 0. Hence (33.22). ∎

Additional useful facts can be established by similar arguments. If $A \subset B$, then

$$(33.23) \quad P[B - A\|\mathscr{G}] = P[B\|\mathscr{G}] - P[A\|\mathscr{G}], \qquad P[A\|\mathscr{G}] \leq P[B\|\mathscr{G}].$$

The inclusion–exclusion formula

$$(33.24) \qquad P\left[\bigcup_{i=1}^{n} A_i\|\mathscr{G}\right] = \sum_i P[A_i\|\mathscr{G}] - \sum_{i<j} P[A_i \cap A_j\|\mathscr{G}] + \cdots$$

holds. If $A_n \uparrow A$, then

$$(33.25) \qquad\qquad\qquad P[A_n\|\mathscr{G}] \uparrow P[A\|\mathscr{G}],$$

and if $A_n \downarrow A$, then

$$(33.26) \qquad\qquad\qquad P[A_n\|\mathscr{G}] \downarrow P[A\|\mathscr{G}].$$

Further, $P(A) = 1$ implies that

$$(33.27) \qquad\qquad\qquad P[A\|\mathscr{G}] = 1,$$

and $P(A) = 0$ implies that

$$(33.28) \qquad\qquad\qquad P[A\|\mathscr{G}] = 0.$$

Of course (33.23) through (33.28) hold with probability 1 only.

Difficulties and Curiosities

This section has been devoted almost entirely to examples connecting the abstract definition (33.8) with the probabilistic idea lying back of it. There are pathological examples showing that the interpretation of conditional probability in terms of an observer with partial information breaks down in certain cases.

Example 33.11. Let $(\Omega, \mathscr{F}, P)$ be the unit interval Ω with Lebesgue measure P on the σ-field $\mathscr{F}$ of Borel subsets of Ω. Take $\mathscr{G}$ to be the σ-field of sets that are either countable or cocountable. Then the function identically equal to $P(A)$ is a version of $P[A\|\mathscr{G}]$: (33.8) holds because $P(G)$ is either 0 or 1 for every G in $\mathscr{G}$. Therefore,

$$(33.29) \qquad\qquad\qquad P[A\|\mathscr{G}]_\omega = P(A)$$

with probability 1. But since $\mathscr{G}$ contains all one-point sets, to know which

elements of $\mathcal{G}$ contain ω is to know ω itself. Thus $\mathcal{G}$ viewed as an experiment should be completely informative—the observer given the information in $\mathcal{G}$ should know ω exactly—and so one might expect that

$$(33.30) \qquad P[A\|\mathcal{G}]_\omega = \begin{cases} 1 & \text{if } \omega \in A, \\ 0 & \text{if } \omega \notin A. \end{cases}$$

This is Example 4.10 in a new form. ∎

The mathematical definition gives (33.29); the heuristic considerations lead to (33.30). Of course, (33.29) is right and (33.30) is wrong. The heuristic view breaks down in certain cases but is nonetheless illuminating and cannot, since it does not intervene in proofs, lead to any difficulties.

The point of view in this section has been "global." To each fixed A in $\mathcal{F}$ has been attached a function (usually a family of functions) $P[A\|\mathcal{G}]_\omega$ defined over all of Ω. What happens if the point of view is reversed—if ω is fixed and A varies over $\mathcal{F}$? Will this result in a probability measure on $\mathcal{F}$? Intuition says it should, and if it does, then (33.21) through (33.28) all reduce to standard facts about measures.

Suppose that $B_1, \ldots, B_r$ is a partition of Ω into $\mathcal{F}$-sets, and let $\mathcal{G} = \sigma(B_1, \ldots, B_r)$. If $P(B_1) = 0$ and $P(B_i) > 0$ for the other i, then one version of $P[A\|\mathcal{G}]$ is

$$P[A\|\mathcal{G}]_\omega = \begin{cases} 17 & \text{if } \omega \in B_1, \\ \dfrac{P(A \cap B_i)}{P(B_i)} & \text{if } \omega \in B_i, \ i = 2, \ldots, r. \end{cases}$$

With this choice of version for each A, $P[A\|\mathcal{G}]_\omega$ is, as a function of A, a probability measure on $\mathcal{F}$ if $\omega \in B_2 \cup \cdots \cup B_r$, but not if $\omega \in B_1$. The "wrong" versions have been chosen. If, for example,

$$P[A\|\mathcal{G}]_\omega = \begin{cases} P(A) & \text{if } \omega \in B_1, \\ \dfrac{P(A \cap B_i)}{P(B_i)} & \text{if } \omega \in B_i, \ i = 2, \ldots, r, \end{cases}$$

then $P[A\|\mathcal{G}]_\omega$ is a probability measure in A for each ω. Clearly, versions such as this one exist if $\mathcal{G}$ is finite.

It might be thought that for an arbitrary σ-field $\mathcal{G}$ in $\mathcal{F}$ versions of the various $P[A\|\mathcal{G}]$ can be so chosen that $P[A\|\mathcal{G}]_\omega$ is for each fixed ω a probability measure as A varies over $\mathcal{F}$. It is possible to construct a

counterexample showing that this is not so.[†] The example is possible because the exceptional ω-set of probability 0 where (33.22) fails depends on the sequence $A_1, A_2, \ldots$; if there are uncountably many such sequences, it can happen that the union of these exceptional sets has positive probability whatever versions $P[A\|\mathscr{G}]$ are chosen.

The existence of such pathological examples turns out not to matter. Example 33.9 illustrates the reason why. From the assumption (33.18) the notably stronger conclusion (33.20) was reached. Since the set $[X_u \in H]$ is fixed throughout the argument, it does not matter that conditional probabilities may not, in fact, be measures. What does matter for the theory is Theorem 33.2 and its extensions.

Consider a point ω_0 with the property that $P(G) > 0$ for every G in $\mathscr{G}$ that contains ω_0. This will be true if the one-point set $\{\omega_0\}$ lies in $\mathscr{F}$ and has positive probability. Fix any versions of the $P[A\|\mathscr{G}]$. For each A the set $[\omega: P[A\|\mathscr{G}]_\omega < 0]$ lies in $\mathscr{G}$ and has probability 0; it therefore cannot contain ω_0. Thus $P[A\|\mathscr{G}]_{\omega_0} \geq 0$. Similarly, $P[\Omega\|\mathscr{G}]_{\omega_0} = 1$, and, if the A_n are disjoint, $P[\bigcup_n A_n\|\mathscr{G}]_{\omega_0} = \sum_n P[A\|\mathscr{G}]_{\omega_0}$. Therefore, $P[A\|\mathscr{G}]_{\omega_0}$ is a probability measure as A ranges over $\mathscr{F}$.

Thus conditional probabilities behave like probabilities at points of positive probability. That they may not do so at points of probability 0 causes no problem because individual such points have no effect on the probabilities of sets. Of course, *sets* of points individually having probability 0 do have an effect, but here the global point of view reenters.

Conditional Probability Distributions

Let X be a random variable on $(\Omega, \mathscr{F}, P)$, and let $\mathscr{G}$ be a σ-field in $\mathscr{F}$.

Theorem 33.3. *There exists a function $\mu(H, \omega)$, defined for H in $\mathscr{R}^1$ and ω in Ω, with these two properties:*

(i) *For each ω in Ω, $\mu(\cdot, \omega)$ is a probability measure on $\mathscr{R}^1$.*
(ii) *For each H in $\mathscr{R}^1$, $\mu(H, \cdot)$ is a version of $P[X \in H\|\mathscr{G}]$.*

The probability measure $\mu(\cdot, \omega)$ is a *conditional distribution* of X given $\mathscr{G}$. If $\mathscr{G} = \sigma(Z)$, it is a conditional distribution of X given Z.

PROOF. For each rational r, let $F(r, \omega)$ be a version of $P[X \leq r\|\mathscr{G}]_\omega$. If $r \leq s$, then by (33.23),

$$(33.31) \qquad\qquad F(r, \omega) \leq F(s, \omega)$$

[†]The argument is outlined in Problem 33.11. It depends on the construction of certain nonmeasurable sets.

for ω outside a $\mathscr{I}$-set A_{rs} of probability 0. By (33.26),

$$(33.32) \qquad\qquad F(r,\omega) = \lim_n F(r + n^{-1}, \omega)$$

for ω outside a $\mathscr{I}$-set B_r of probability 0. Finally, by (33.25) and (33.26),

$$(33.33) \qquad \lim_{r \to -\infty} F(r,\omega) = 0, \qquad \lim_{r \to \infty} F(r,\omega) = 1$$

outside a $\mathscr{I}$-set C of probability 0. As there are only countably many of these exceptional sets, their union E lies in $\mathscr{I}$ and has probability 0.

For $\omega \notin E$ extend $F(\cdot, \omega)$ to all of R^1 by setting $F(x,\omega) = \inf[F(r,\omega): x < r]$. For $\omega \in E$ take $F(x,\omega) = F(x)$, where F is some arbitrary but fixed distribution function. Suppose that $\omega \notin E$. By (33.31) and (33.32), $F(x,\omega)$ agrees with the first definition on the rationals and is nondecreasing; it is right-continuous; and by (33.33) it is a probability distribution function. Therefore, there exists a probability measure $\mu(\cdot, \omega)$ on $(R^1, \mathscr{R}^1)$ with distribution function $F(\cdot, \omega)$. For $\omega \in E$, let $\mu(\cdot, \omega)$ be the probability measure corresponding to $F(x)$. Then condition (i) is satisfied.

The class of H for which $\mu(H, \cdot)$ is measurable $\mathscr{I}$ is a λ-system containing the sets $H = (-\infty, r]$ for rational r; therefore $\mu(H, \cdot)$ is measurable $\mathscr{I}$ for H in $\mathscr{R}^1$.

By construction, $\mu((-\infty, r], \omega) = P[X \le r \| \mathscr{I}]_\omega$ with probability 1 for rational r; that is, for $H = (-\infty, r]$ as well as for $H = R^1$,

$$\int_G \mu(H, \omega) P(d\omega) = P([X \in H] \cap G)$$

for all G in $\mathscr{I}$. Fix G. Each side of this equation is a measure as a function of H, and so the equation must hold for all H in $\mathscr{R}^1$. ∎

Example 33.12. Let X and Y be random variables whose joint distribution ν in R^2 has density $f(x, y)$ with respect to Lebesgue measure: $P[(X, Y) \in A] = \nu(A) = \iint_A f(x, y)\, dx\, dy$. Let $g(x, y) = f(x, y)/\int_{R^1} f(x, t)\, dt$, and let $\mu(H, x) = \int_H g(x, y)\, dy$ have probability density $g(x, \cdot)$; if $\int_{R^1} f(x, t)\, dt = 0$, let $\mu(\cdot, x)$ be an arbitrary probability measure on the line. Then $\mu(H, X(\omega))$ will serve as the conditional distribution of Y given X. Indeed, (33.10) is the same thing as $\int_{E \times R^1} \mu(F, x)\, d\nu(x, y) = \nu(E \times F)$, and a change of variable gives $\int_{[X \in E]} \mu(F, X(\omega)) P(d\omega) = P[X \in E, Y \in F]$. Thus $\mu(F, X(\omega))$ is a version of $P[Y \in F \| X]_\omega$. This is a new version of Example 33.5. ◨

PROBLEMS

33.1. 20.27↑ *Borel's paradox.* Suppose that a random point on the sphere is specified by longitude Θ and latitude Φ, but restrict Θ by $0 \le \Theta < \pi$, so that Θ specifies the complete meridian circle (not semicircle) containing the point, and compensate by letting Φ range over $(-\pi, \pi]$.

(a) Show that for given Θ the conditional distribution of Φ has density $\frac{1}{4}|\cos \phi|$ over $(-\pi, +\pi]$. If the point lies on, say, the meridian circle through Greenwich, it is therefore not uniformly distributed over that great circle.

(b) Show that for given Φ the conditional distribution of Θ is uniform over $(0, \pi)$. If the point lies on the equator (Φ is 0 or π), it is therefore uniformly distributed over that great circle.

Since the point is uniformly distributed over the spherical surface and great circles are indistinguishable, (a) and (b) stand in apparent contradiction. This shows again the inadmissibility of conditioning with respect to an isolated event of probability 0. The relevant σ-field must not be lost sight of.

33.2. 20.16↑ Let X and Y be independent, each having the standard normal distribution, and let (R, Θ) be the polar coordinates for (X, Y).

(a) Show that $X + Y$ and $X - Y$ are independent and that $R^2 = [(X + Y)^2 + (X - Y)^2]/2$, and conclude that the conditional distribution of R^2 given $X - Y$ is the chi-squared distribution with one degree of freedom translated by $(X - Y)^2/2$.

(b) Show that the conditional distribution of R^2 given Θ is chi-squared with two degrees of freedom.

(c) If $X - Y = 0$, the conditional distribution of R^2 is chi-squared with one degree of freedom. If $\Theta = \pi/4$ or $\Theta = 5\pi/4$, the conditional distribution of R^2 is chi-squared with two degrees of freedom. But the events $[X - Y = 0]$ and $[\Theta = \pi/4] \cup [\Theta = 5\pi/4]$ are the same. Resolve the apparent contradiction.

33.3. ↑ Paradoxes of a somewhat similar kind arise in very simple cases.

(a) Of three prisoners, call them 1, 2, and 3, two have been chosen by lot for execution. Prisoner 3 says to the guard, "Which of 1 and 2 is to be executed? One of them will be, and you give me no information about myself in telling me which it is." The guard finds this reasonable and says, "Prisoner 1 is to be executed." And now 3 reasons, "I know that 1 is to be executed; the other will be either 2 or me, and so my chance of being executed is now only $\frac{1}{2}$, instead of the $\frac{2}{3}$ it was before," Apparently, the guard *has* given him information.

If one looks for a σ-field, it must be the one describing the guard's answer, and it then becomes clear that the sample space is incompletely specified. Suppose that, if 1 and 2 are to be executed, the guard's response is "1" with probability p and "2" with probability $1 - p$; and, of course, suppose that, if 3 is to be executed, the guard names the other victim. Calculate the conditional probabilities.

(b) Assume that among families with two children the four sex distributions are equally likely. You have been introduced to one of the two children in such a family, and he is a boy. What is the conditional probability that the other is a boy as well?

33.4. (a) Consider probability spaces $(\Omega, \mathscr{F}, P)$ and $(\Omega', \mathscr{F}', P')$; suppose that T: $\Omega \to \Omega'$ is measurable $\mathscr{F}/\mathscr{F}'$ and $P' = PT^{-1}$. Let $\mathscr{G}'$ be a σ-field in $\mathscr{F}'$, and take $\mathscr{G}$ to be the σ-field $[T^{-1}G': G' \in \mathscr{G}']$. For $A' \in \mathscr{F}'$, show by (16.18) that $P[T^{-1}A'\|\mathscr{G}]_\omega = P'[A'\|\mathscr{G}']_{T\omega}$ with P-probability 1.

(b) Now take $(\Omega', \mathscr{F}', P') = (R^2, \mathscr{R}^2, \mu)$, where μ is the distribution of a random vector (X, Y) on $(\Omega, \mathscr{F}, P)$. Suppose that (X, Y) has density f, and show by (33.9) that

$$P[Y \in F\| X]_\omega = \frac{\int_F f(X(\omega), t)\, dt}{\int_{R^1} f(X(\omega), t)\, dt}$$

with probability 1.

33.5. ↑ (a) There is a slightly different approach to conditional probability. Let $(\Omega, \mathscr{F}, P)$ be a probability space, $(\Omega', \mathscr{F}')$ a measurable space, and T: $\Omega \to \Omega'$ a mapping measurable $\mathscr{F}/\mathscr{F}'$. Define a measure ν on $\mathscr{F}'$ by $\nu(A') = P(A \cap T^{-1}A')$ for $A' \in \mathscr{F}'$. Prove that there exists a function $p(A|\omega')$ on Ω', measurable $\mathscr{F}'$ and integrable PT^{-1}, such that $\int_{A'} p(A|\omega')PT^{-1}(d\omega') = P(A \cap T^{-1}A')$ for all A' in $\mathscr{F}'$. Intuitively, $p(A|\omega')$ is the conditional probability that $\omega \in A$ for someone who knows that $T\omega = \omega'$. Let $\mathscr{G} = [T^{-1}A': A' \in \mathscr{F}']$; show that $\mathscr{G}$ is a σ-field and that $p(A|T\omega)$ is a version of $P[A\|\mathscr{G}]_\omega$.

(b) Connect this with part (a) of the preceding problem.

33.6. ↑ Suppose that $T = X$ is a random variable, $(\Omega', \mathscr{F}') = (R^1, \mathscr{R}^1)$, and x is the general point of R^1. In this case $p(A|x)$ is sometimes written $P[A|X = x]$. What is the problem with this notation?

33.7. For the Poisson process (see Example 33.1) show that for $0 < s < t$,

$$P[N_s = k\|N_t] = \begin{cases} \binom{N_t}{k}\left(\frac{s}{t}\right)^k\left(1 - \frac{s}{t}\right)^{N_t - k}, & k \le N_t, \\ 0, & k > N_t. \end{cases}$$

Thus the conditional distribution (in the sense of Theorem 33.3) of N_s given N_t is binomial with parameters N_t and s/t.

33.8. 29.12↑ Suppose that (X_1, X_2) has the centered normal distribution—has in the plane the distribution with density (29.10). Express the quadratic form in the exponential as

$$\frac{1}{\sigma_{11}}x_1^2 + \frac{\sigma_{11}}{D}\left(x_2 - \frac{\sigma_{12}}{\sigma_{11}}x_1\right)^2;$$

integrate out the x_2 and show that

$$\frac{f(x_1, x_2)}{\int_{-\infty}^{\infty} f(x_1, t)\, dt} = \frac{1}{\sqrt{2\pi\tau}} \exp\left[-\frac{1}{2\tau}\left(x_2 - \frac{\sigma_{12}}{\sigma_{11}} x_1 \right)^2 \right],$$

where $\tau = \sigma_{22} - \sigma_{12}^2 \sigma_{11}^{-1}$. Describe the conditional distribution of X_2 given X_1.

33.9. (a) Suppose that $\mu(H, \omega)$ has property (i) in Theorem 33.3, and suppose that $\mu(H, \cdot)$ is a version of $P[X \in H \| \mathscr{S}]$ for H in a π-system generating $\mathscr{R}^1$. Show that $\mu(\cdot, \omega)$ is a conditional distribution of X given $\mathscr{S}$.

(b) Use Theorem 12.5 to extend Theorem 33.3 from R^1 to R^k.

(c) Show that conditional probabilities can be defined as genuine probabilities on spaces of the special form $(\Omega, \sigma(X_1, \ldots, X_k), P)$.

33.10. ↑ Deduce from (33.16) that the conditional distribution of X given M is

$$\frac{1}{2} I_{[M \in H]}(\omega) + \frac{1}{2} \frac{\mu(H \cap (-\infty, M(\omega)])}{\mu(-\infty, M(\omega)])},$$

where μ is the distribution corresponding to F (positive and continuous). *Hint:* First check $H = (-\infty, x]$.

33.11. 4.10   12.4↑   The following construction shows that conditional probabilities may not give measures. Complete the details.

In Problem 4.10 it is shown that there exist a probability space $(\Omega, \mathscr{F}, P)$, a σ-field $\mathscr{S}$ in $\mathscr{F}$, and a set H in $\mathscr{F}$ such that $P(H) = \frac{1}{2}$, H and $\mathscr{S}$ are independent, $\mathscr{S}$ contains all the singletons, and $\mathscr{S}$ is generated by a countable subclass. The countable subclass generating $\mathscr{S}$ can be taken to be a π-system $\mathscr{P} = \{B_1, B_2, \ldots\}$ (pass to the finite intersections of the sets in the original class).

Assume that it is possible to choose versions $P[A\|\mathscr{S}]$ so that $P[A\|\mathscr{S}]_\omega$ is for each ω a probability measure as A varies over $\mathscr{F}$. Let C_n be the ω-set where $P[B_n\|\mathscr{S}]_\omega = I_{B_n}(\omega)$; show (Example 33.3) that $C = \bigcap_n C_n$ has probability 1. Show that $\omega \in C$ implies that $P[G\|\mathscr{S}]_\omega = I_G(\omega)$ for all G in $\mathscr{S}$ and hence that $P[\{\omega\}\|\mathscr{S}]_\omega = 1$.

Now $\omega \in H \cap C$ implies that $P[H\|\mathscr{S}]_\omega \geq P[\{\omega\}\|\mathscr{S}]_\omega = 1$ and $\omega \in H^c \cap C$ implies that $P[H\|\mathscr{S}]_\omega \leq P[\Omega - \{\omega\}\|\mathscr{S}]_\omega = 0$. Thus $\omega \in C$ implies that $P[H\|\mathscr{S}]_\omega = I_H(\omega)$. But since H and $\mathscr{S}$ are independent, $P[H\|\mathscr{S}]_\omega = P(H) = \frac{1}{2}$ with probability 1, a contradiction.

This example is related to Example 4.10 but concerns mathematical fact instead of heuristic interpretation.

33.12. Let α and β be σ-finite measures on the line, and let $f(x, y)$ be a probability density with respect to $\alpha \times \beta$. Define

$$(33.34) \qquad\qquad g_x(y) = \frac{f(x, y)}{\int_{R^1} f(x, t)\beta(dt)},$$

unless the denominator vanishes, in which case take $g_x(y) = 0$, say. Show that, if (X, Y) has density f with respect to $\alpha \times \beta$, then the conditional distribution of Y given X has density $g_X(y)$ with respect to β. This generalizes Examples 33.5 and 33.12, where α and β are Lebesgue measure.

33.13. 18.20↑ Suppose that μ and ν_x (one for each real x) are probability measures on the line, and suppose that $\nu_x(B)$ is a Borel function in x for each $B \in \mathscr{R}^1$. Then (see Problem 18 20)

$$(33.35) \qquad \pi(E) = \int_{R^1} \nu_x[\, y \cdot (x, y) \in E] \mu(dx)$$

defines a probability measure on $(R^2, \mathscr{R}^2)$.

Suppose that (X, Y) has distribution π, and show that ν_X is a version of the conditional distribution of Y given X.

33.14. ↑ Let α and β be σ-finite measures on the line. Specialize the setup of Problem 33.13 by supposing that μ has density $f(x)$ with respect to α and ν_x has density $g_x(y)$ with respect to β. Assume that $g_x(y)$ is measurable $\mathscr{R}^2$ in the pair (x, y), so that $\nu_x(B)$ is automatically measurable in x. Show that (33.35) has density $f(x)g_x(y)$ with respect to $\alpha \times \beta$: $\pi(E) = \int\int_E f(x)g_x(y)\alpha(dx)\beta(dy)$. Show that (33.34) is consistent with $f(x, y) = f(x)g_x(y)$. Put

$$p_y(x) = \frac{f(x)g_x(y)}{\int_{R^1} f(s)g_s(y)\alpha(ds)}.$$

Suppose that (X, Y) has density $f(x)g_x(y)$ with respect to $\alpha \times \beta$, and show that $p_Y(x)$ is a density with respect to α for the conditional distribution of X given Y.

In the language of Bayes, $f(x)$ is the prior density of a parameter x, $g_x(y)$ is the conditional density of the observation y given the parameter, and $p_y(x)$ is the posterior density of the parameter given the observation.

33.15. ↑ Now suppose that α and β are Lebesgue measure, that $f(x)$ is positive, continuous, and bounded, and that $g_x(y) = e^{-(y-x)^2 n/2}/\sqrt{2\pi/n}$. Thus the observation is distributed as the average of n independent normal variables with mean x and variance 1. Show that

$$\frac{1}{\sqrt{n}} p_y\left(y + \frac{x}{\sqrt{n}}\right) \to \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

for fixed x and y. Thus the posterior density is approximately that of a normal distribution with mean y and variance $1/n$.

33.16. 32.13↑ Suppose that X has distribution μ. Now $P[A\|X]_\omega = f(X(\omega))$ for some Borel function f. Show that $\lim_{h \to 0} P[A|x - h < X \le x + h] = f(x)$ for x in a set of μ-measure 1. Roughly speaking, $P[A|x - h < X \le x + h] \to P[A|X = x]$. *Hint:* Take $\nu(B) = P(A \cap [X \in B])$ in Problem 32.13.

SECTION 34. CONDITIONAL EXPECTATION

In this section the theory of conditional expectation is developed from first principles. The properties of conditional probabilities will then follow as special cases. The preceding section was long only because of the examples in it; the theory itself is quite compact.

Definition

Suppose that X is an integrable random variable on $(\Omega, \mathscr{F}, P)$ and that $\mathscr{G}$ is a σ-field in $\mathscr{F}$. There exists a random variable $E[X\|\mathscr{G}]$, called the *conditional expected value* of X given $\mathscr{G}$, having these two properties:

(i) $E[X\|\mathscr{G}]$ *is measurable $\mathscr{G}$ and integrable.*
(ii) $E[X\|\mathscr{G}]$ *satisfies the functional equation*

$$(34.1) \qquad \int_G E[X\|\mathscr{G}]\, dP = \int_G X\, dP, \qquad G \in \mathscr{G}.$$

To prove the existence of such a random variable, consider first the case of nonnegative X. Define a measure ν on $\mathscr{G}$ by $\nu(G) = \int_G X\, dP$. This measure is finite because X is integrable, and it is absolutely continuous with respect to P. By the Radon–Nikodym theorem there is a function f, measurable $\mathscr{G}$, such that $\nu(G) = \int_G f\, dP$.[†] This f has properties (i) and (ii). If X is not necessarily nonnegative, $E[X^+\|\mathscr{G}] - E[X^-\|\mathscr{G}]$ clearly has the required properties.

There will in general be many such random variables $E[X\|\mathscr{G}]$; any one of them is called a *version* of the conditional expected value. Any two versions are equal with probability 1 (by Theorem 16.10 applied to P restricted to $\mathscr{G}$).

Arguments like those in Examples 33.3 and 33.4 show that $E[X\|\{0, \Omega\}] = E[X]$ and that $E[X\|\mathscr{F}] = X$ with probability 1. As $\mathscr{G}$ increases, condition (i) becomes weaker and condition (ii) becomes stronger.

The value $E[X\|\mathscr{G}]_\omega$ at ω is to be interpreted as the expected value of X for someone who knows for each G in $\mathscr{G}$ whether or not it contains the point ω, which itself in general remains unknown. Condition (i) ensures that $E[X\|\mathscr{G}]$ can in principle be calculated from this partial information alone. Condition (ii) can be restated as $\int_G (X - E[X\|\mathscr{G}])\, dP = 0$; if the observer, in possession of the partial information contained in $\mathscr{G}$, is offered the opportunity to bet, paying an entry fee of $E[X\|\mathscr{G}]$ and being returned the amount X, and if he adopts the strategy of betting if G occurs, this equation says that the game is fair.

[†]As in the case of conditional probabilities, the integral is the same on $(\Omega, \mathscr{F}, P)$ as on $(\Omega, \mathscr{G})$ with P restricted to $\mathscr{G}$ (Example 16.4).

Example 34.1. Suppose that $B_1, B_2, \ldots$ is a finite or countable partition of Ω generating the σ-field $\mathscr{G}$. Then $E[X\|\mathscr{G}]$ must, since it is measurable $\mathscr{G}$, have some constant value over B_i, say α_i. Then (34.1) for $G = B_i$ gives $\alpha_i P(B_i) = \int_{B_i} X \, dP$. Thus

$$(34.2) \qquad E[X\|\mathscr{G}]_\omega = \frac{1}{P(B_i)} \int_{B_i} X \, dP, \qquad \omega \in B_i, \qquad P(B_i) > 0.$$

If $P(B_i) = 0$, the value of $E[X\|\mathscr{G}]$ over B_i is constant but arbitrary.   ∎

Example 34.2. For an indicator I_A the defining properties of $E[I_A\|\mathscr{G}]$ and $P[A\|\mathscr{G}]$ coincide; therefore, $E[I_A\|\mathscr{G}] = P[A\|\mathscr{G}]$ with probability 1. It is easily checked that, more generally, $E[X\|\mathscr{G}] = \sum_i \alpha_i P[A_i\|\mathscr{G}]$ with probability 1 for a simple function $X = \sum_i \alpha_i I_{A_i}$.   ∎

In analogy with the case of conditional probability, if $[X_t, \ t \in T]$ is a collection of random variables, $E[X\|X_t, \ t \in T]$ is by definition $E[X\|\mathscr{G}]$ with $\sigma[X_t, \ t \in T]$ in the role of $\mathscr{G}$.

Example 34.3. Let $\mathscr{I}$ be the σ-field of sets invariant under a measure-preserving transformation T on $(\Omega, \mathscr{F}, P)$. For f integrable, the limit $\hat{f}$ in (24.7) is $E[f\|\mathscr{I}]$: Since $\hat{f}$ is invariant, it is measurable $\mathscr{I}$. If G is invariant, then the averages a_n in the proof of the ergodic theorem (p. 318) satisfy $E[I_G a_n] = E[I_G f]$. But since the a_n converge to $\hat{f}$ and are uniformly integrable, $E[I_G \hat{f}] = E[I_G f]$.   ∎

Properties of Conditional Expectation

To prove the first result, apply Theorem 16.10(iii) to f and $E[X\|\mathscr{G}]$ on $(\Omega, \mathscr{G}, P)$.

Theorem 34.1. *Let $\mathscr{P}$ be a π-system generating the σ-field $\mathscr{G}$, and suppose that Ω is a finite or countable union of sets in $\mathscr{P}$. An integrable function f is a version of $E[X\|\mathscr{G}]$ if it is measurable $\mathscr{G}$ and if*

$$(34.3) \qquad \int_G f \, dP = \int_G X \, dP$$

holds for all G in $\mathscr{P}$.

In most applications it is clear that $\Omega \in \mathscr{P}$.

All the equalities and inequalities in the following theorem hold with probability 1.

Theorem 34.2. *Suppose that* X, Y, X_n *are integrable.*

(i) *If* $X = a$ *with probability* 1, *then* $E[X\|\mathscr{G}] = a$.

(ii) *For constants* a *and* b, $E[aX + bY\|\mathscr{G}] = aE[X\|\mathscr{G}] + bE[Y\|\mathscr{G}]$.

(iii) *If* $X \le Y$ *with probability* 1, *then* $E[X\|\mathscr{G}] \le E[Y\|\mathscr{G}]$.

(iv) $|E[X\|\mathscr{G}]| \le E[|X|\|\mathscr{G}]$.

(v) *If* $\lim_n X_n = X$ *with probability* 1, $|X_n| \le Y$, *and* Y *is integrable, then* $\lim_n E[X_n\|\mathscr{G}] = E[X\|\mathscr{G}]$ *with probability* 1.

PROOF. If $X = a$ with probability 1, the function identically equal to a satisfies conditions (i) and (ii) in the definition of $E[X\|\mathscr{G}]$, and so (i) above follows by uniqueness.

As for (ii), $aE[X\|\mathscr{G}] + bE[Y\|\mathscr{G}]$ is integrable and measurable $\mathscr{G}$, and

$$\int_G (aE[X\|\mathscr{G}] + bE[Y\|\mathscr{G}])\, dP = a\int_G E[X\|\mathscr{G}]\, dP + b\int_G E[Y\|\mathscr{G}]\, dP$$

$$= a\int_G X\, dP + b\int_G Y\, dP = \int_G (aX + bY)\, dP$$

for all G in $\mathscr{G}$, so that this function satisfies the functional equation.

If $X \le Y$ with probability 1, then $\int_G (E[Y\|\mathscr{G}] - E[X\|\mathscr{G}])\, dP = \int_G (Y - X)\, dP \ge 0$ for all G in $\mathscr{G}$. Since $E[Y\|\mathscr{G}] - E[X\|\mathscr{G}]$ is measurable $\mathscr{G}$, it must be nonnegative with probability 1 (consider the set G where it is negative). This proves (iii), which clearly implies (iv) as well as the fact that $E[X\|\mathscr{G}] = E[Y\|\mathscr{G}]$ if $X = Y$ with probability 1.

To prove (v), consider $Z_n = \sup_{k \ge n}|X_k - X|$. Now $Z_n \downarrow 0$ with probability 1, and by (ii), (iii), and (iv), $|E[X_n\|\mathscr{G}] - E[X\|\mathscr{G}]| \le E[Z_n\|\mathscr{G}]$. It suffices, therefore, to show that $E[Z_n\|\mathscr{G}] \downarrow 0$ with probability 1. By (iii) the sequence $E[Z_n\|\mathscr{G}]$ is nonincreasing and hence has a limit Z; the problem is to prove that $Z = 0$ with probability 1, or, Z being nonnegative, that $E[Z] = 0$. But $0 \le Z_n \le 2Y$, and so (34.1) and the dominated convergence theorem give $E[Z] = \int E[Z\|\mathscr{G}]\, dP \le \int E[Z_n\|\mathscr{G}]\, dP = E[Z_n] \to 0$. ∎

The properties (33.21) through (33.28) can be derived anew from Theorem 34.2. Part (ii) shows once again that $E[\sum_i \alpha_i I_{A_i}\|\mathscr{G}] = \sum_i \alpha_i P[A_i\|\mathscr{G}]$ for simple functions.

If X is measurable $\mathscr{G}$, then clearly $E[X\|\mathscr{G}] = X$ with probability 1. The following generalization of this is used constantly. For an observer with the information in $\mathscr{G}$, X is effectively a constant if it is measurable $\mathscr{G}$:

Theorem 34.3. *If* X *is measurable* $\mathscr{G}$, *and if* Y *and* XY *are integrable, then*

$$(34.4) \qquad\qquad E[XY\|\mathscr{G}] = XE[Y\|\mathscr{G}]$$

with probability 1.

PROOF. It will be shown first that the right side of (34.4) is a version of the left side if $X = I_{G_0}$ and $G_0 \in \mathscr{G}$. Since $I_{G_0} E[Y \| \mathscr{G}]$ is certainly measurable $\mathscr{G}$, it suffices to show that it satisfies the functional equation $\int_G I_{G_0} E[Y \| \mathscr{G}] dP = \int_G I_{G_0} Y dP$, $G \in \mathscr{G}$. But this reduces to $\int_{G \cap G_0} E[Y \| \mathscr{G}] dP = \int_{G \cap G_0} Y dP$, which holds by the definition of $E[Y \| \mathscr{G}]$. Thus (34.4) holds if X is the indicator of an element of $\mathscr{G}$.

It follows by Theorem 34.2(ii) that (34.4) holds if X is a simple function measurable $\mathscr{G}$. For the general X that is measurable $\mathscr{G}$, there exist simple functions X_n, measurable $\mathscr{G}$, such that $|X_n| \leq |X|$ and $\lim_n X_n = X$ (Theorem 13.5). Since $|X_n Y| \leq |XY|$ and $|XY|$ is integrable, Theorem 34.2(v) implies that $\lim_n E[X_n Y \| \mathscr{G}] = E[XY \| \mathscr{G}]$ with probability 1. But $E[X_n Y \| \mathscr{G}] = X_n E[Y \| \mathscr{G}]$ by the case already treated, and of course $\lim_n X_n E[Y \| \mathscr{G}] = XE[Y \| \mathscr{G}]$. (Note that $|X_n E[Y \| \mathscr{G}]| = |E[X_n Y \| \mathscr{G}]| \leq E[|X_n Y| \| \mathscr{G}] \leq E[|XY| \| \mathscr{G}]$, so that the limit $XE[Y \| \mathscr{G}]$ is integrable.) Thus (34.4) holds in general. Notice that X has not been assumed integrable. ∎

Taking a conditional expected value can be thought of as an averaging or smoothing operation. This leads one to expect that averaging X with respect to $\mathscr{G}_2$ and then averaging the result with respect to a coarser (smaller) σ-field $\mathscr{G}_1$ should lead to the same result as would averaging with respect to $\mathscr{G}_1$ in the first place:

Theorem 34.4. *If X is integrable and the σ-fields $\mathscr{G}_1$ and $\mathscr{G}_2$ satisfy $\mathscr{G}_1 \subset \mathscr{G}_2$, then*

$$(34.5) \qquad E\big[E[X \| \mathscr{G}_2] \| \mathscr{G}_1\big] = E[X \| \mathscr{G}_1]$$

with probability 1.

PROOF. The left side of (34.5) is measurable $\mathscr{G}_1$, and so to prove that it is a version of $E[X \| \mathscr{G}_1]$, it is enough to verify $\int_G E[E[X \| \mathscr{G}_2] \| \mathscr{G}_1] dP = \int_G X dP$ for $G \in \mathscr{G}_1$. But if $G \in \mathscr{G}_1$, then $G \in \mathscr{G}_2$, and the left side here is $\int_G E[X \| \mathscr{G}_2] dP = \int_G X dP$. ∎

If $\mathscr{G}_2 = \mathscr{F}$, then $E[X \| \mathscr{G}_2] = X$, so that (34.5) is trivial. If $\mathscr{G}_1 = \{0, \Omega\}$ and $\mathscr{G}_2 = \mathscr{G}$, then (34.5) becomes

$$(34.6) \qquad E[E[X \| \mathscr{G}]] = E[X],$$

the special case of (34.1) for $G = \Omega$.

If $\mathscr{G}_1 \subset \mathscr{G}_2$, then $E[X \| \mathscr{G}_1]$, being measurable $\mathscr{G}_1$, is also measurable $\mathscr{G}_2$, so that taking an expected value with respect to $\mathscr{G}_2$ does not alter it: $E[E[X \| \mathscr{G}_1] \| \mathscr{G}_2] = E[X \| \mathscr{G}_1]$. Therefore, if $\mathscr{G}_1 \subset \mathscr{G}_2$, taking iterated expected values in either order gives $E[X \| \mathscr{G}_1]$.

The remaining result of a general sort needed here is *Jensen's inequality* for conditional expected values: If φ is a convex function on the line and X and $\varphi(X)$ are both integrable, then

$$(34.7) \qquad\qquad \varphi(E[X\|\mathscr{G}]) \le E[\varphi(X)\|\mathscr{G}]$$

with probability 1. For each x_0 take a support line [A33] through $(x_0, \varphi(x_0))$: $\varphi(x_0) + A(x_0)(x - x_0) \le \varphi(x)$. The slope $A(x_0)$ can be taken as the right-hand derivative of φ, so that it is nondecreasing in x_0. Now

$$\varphi(E[X\|\mathscr{G}]) + A(E[X\|\mathscr{G}])(X - E[X\|\mathscr{G}]) \le \varphi(X).$$

Suppose that $E[X\|\mathscr{G}]$ is bounded. Then all three terms here are integrable (if φ is convex on R^1, then φ and A are bounded on bounded sets), and taking expected values with respect to $\mathscr{G}$ and using (34.4) on the middle term gives (34.7).

To prove (34.7) in general, let $G_n = [|E[X\|\mathscr{G}]| \le n]$. Then $E[I_{G_n}X\|\mathscr{G}] = I_{G_n}E[X\|\mathscr{G}]$ is bounded, and so (34.7) holds for $I_{G_n}X$: $\varphi(I_{G_n}E[X\|\mathscr{G}]) \le E[\varphi(I_{G_n}X)\|\mathscr{G}]$. Now $E[\varphi(I_{G_n}X)\|\mathscr{G}] = E[I_{G_n}\varphi(X) + I_{G_n^c}\varphi(0)\|\mathscr{G}] = I_{G_n}E[\varphi(X)\|\mathscr{G}] + I_{G_n^c}\varphi(0) \to E[\varphi(X)\|\mathscr{G}]$. Since $\varphi(I_{G_n}E[X\|\mathscr{G}])$ converges to $\varphi(E[X\|\mathscr{G}])$ by the continuity of φ, (34.7) follows. If $\varphi(x) = |x|$, (34.7) gives part (iv) of Theorem 34.2 again.

Conditional Distributions and Expectations

Theorem 34.5. *Let $\mu(\cdot, \omega)$ be a conditional distribution with respect to $\mathscr{G}$ of a random variable X, in the sense of Theorem 33.3. If φ: $R^1 \to R^1$ is a Borel function for which $\varphi(X)$ is integrable, then $\int_{R^1}\varphi(x)\mu(dx, \omega)$ is a version of $E[\varphi(X)\|\mathscr{G}]_\omega$.*

PROOF. If $\varphi = I_H$ and $H \in \mathscr{R}^1$, this is an immediate consequence of the definition of conditional distribution, and by Theorem 34.2(ii) it follows for φ a simple function over R^1. For the general nonnegative φ, choose simple φ_n such that $0 \le \varphi_n(x) \uparrow \varphi(x)$ for each x in R^1. By the case already treated, $\int_{R^1}\varphi_n(x)\mu(dx, \omega)$ is a version of $E[\varphi_n(X)\|\mathscr{G}]_\omega$. The integral converges by the monotone convergence theorem in $(R^1, \mathscr{R}^1, \mu(\cdot, \omega))$ to $\int_{R^1}\varphi(x)\mu(dx, \omega)$ for each ω, the value $+\infty$ not excluded, and $E[\varphi_n(X)\|\mathscr{G}]_\omega$ converges to $E[\varphi(X)\|\mathscr{G}]_\omega$ with probability 1 by Theorem 34.2(v). Thus the result holds for nonnegative φ, and the general case follows from splitting into positive and negative parts. ∎

It is a consequence of the proof above that $\int_{R^1}\varphi(x)\mu(dx, \omega)$ is measurable $\mathscr{G}$ and finite with probability 1. If X is itself integrable, it follows by the

theorem for the case $\varphi(x) = x$ that

$$E[X\|\mathscr{G}]_\omega = \int_{-\infty}^{\infty} x\mu(dx, \omega)$$

with probability 1. If $\varphi(X)$ is integrable as well, then

$$(34.8) \qquad E[\varphi(X)\|\mathscr{G}]_\omega = \int_{-\infty}^{\infty} \varphi(x)\mu(dx, \omega)$$

with probability 1. By Jensen's inequality (21.14) for unconditional expected values, the right side of (34.8) is at least $\varphi(\int_{-\infty}^{\infty} x\mu(dx, \omega))$ if φ is convex. This gives another proof of (34.7).

Sufficient Subfields*

Suppose that for each θ in an index set Θ, P_θ is a probability measure on $(\Omega, \mathscr{F})$. In statistics the problem is to draw inferences about the unknown parameter θ from an observation ω.

Denote by $P_\theta[A\|\mathscr{G}]$ and $E_\theta[X\|\mathscr{G}]$ conditional probabilities and expected values calculated with respect to the probability measure P_θ on $(\Omega, \mathscr{F})$. A σ-field $\mathscr{G}$ in $\mathscr{F}$ is *sufficient* for the family $[P_\theta: \theta \in \Theta]$ if versions $P_\theta[A\|\mathscr{G}]$ can be chosen that are independent of θ—that is, if there exists a function $p(A, \omega)$ of $A \in \mathscr{F}$ and $\omega \in \Omega$ such that, for each $A \in \mathscr{F}$ and $\theta \in \Theta$, $p(A, \cdot)$ is a version of $P_\theta[A\|\mathscr{G}]$. There is no requirement that $p(\cdot, \omega)$ be a measure for ω fixed. The idea is that although there may be information in $\mathscr{F}$ not already contained in $\mathscr{G}$, this information is irrelevant to the drawing of inferences about θ.[†] A *sufficient statistic* is a random variable or random vector T such that $\sigma(T)$ is a sufficient subfield.

A family $\mathscr{M}$ of measures *dominates* another family $\mathscr{N}$ if, for each A, from $\mu(A) = 0$ for all μ in $\mathscr{M}$, it follows that $\nu(A) = 0$ for all ν in $\mathscr{N}$. If each of $\mathscr{M}$ and $\mathscr{N}$ dominates the other, they are *equivalent*. For sets consisting of a single measure these are the concepts introduced in Section 32.

Theorem 34.6. *Suppose that* $[P_\theta: \theta \in \Theta]$ *is dominated by the σ-finite measure μ. A necessary and sufficient condition for $\mathscr{G}$ to be sufficient is that the density f_θ of P_θ with respect to μ can be put in the form $f_\theta = g_\theta h$ for a g_θ measurable $\mathscr{G}$.*

It is assumed throughout that g_θ and h are nonnegative and of course that h is measurable $\mathscr{F}$. Theorem 34.6 is called the *factorization theorem*, the condition being that the density f_θ splits into a factor depending on ω only through $\mathscr{G}$ and a factor independent of θ. Although g_θ and h are not assumed integrable μ, their product f_θ, as the density of a finite measure, must be. Before proceeding to the proof, consider an application.

Example 34.4. Let $(\Omega, \mathscr{F}) = (R^k, \mathscr{R}^k)$, and for $\theta > 0$ let P_θ be the measure having with respect to k-dimensional Lebesgue measure the density

$$f_\theta(x) = f_\theta(x_1, \ldots, x_k) = \begin{cases} \theta^{-k} & \text{if } 0 \le x_i \le \theta, i = 1, \ldots, k, \\ 0 & \text{otherwise.} \end{cases}$$

---

*This topic may be omitted.
[†]See Problem 34.19.

If X_i is the function on R^k defined by $X_i(x) = x_i$, then under P_θ, $X_1, \ldots, X_k$ are independent random variables, each uniformly distributed over $[0, \theta]$. Let $T(x) = \max_{i \le k} X_i(x)$. If $g_\theta(t)$ is θ^{-k} for $0 \le t \le \theta$ and 0 otherwise, and if $h(x)$ is 1 or 0 according as all x_i are nonnegative or not, then $f_\theta(x) = g_\theta(T(x))h(x)$. The factorization criterion is thus satisfied, and T is a sufficient statistic.

Sufficiency is clear on intuitive grounds as well: θ is not involved in the conditional distribution of $X_1, \ldots, X_k$ given T because, roughly speaking, a random one of them equals T and the others are independent and uniform over $[0, T]$. If this is true, the distribution of X_i given T ought to have a mass of k^{-1} at T and a uniform distribution of mass $1 - k^{-1}$ over $[0, T]$, so that

$$(34.9) \qquad E_\theta[X_i \| T] = \frac{1}{k}T + \frac{k-1}{k}\frac{T}{2} = \frac{k+1}{2k}T.$$

For a proof of this fact, needed later, note that by (21 9)

$$(34\ 10) \qquad \int_{T \le t} X_i\, dP_\theta = \int_0^\infty P_\theta[T \le t,\ X_i \ge u]\, du$$

$$= \int_0^t \frac{t-u}{\theta}\left(\frac{t}{\theta}\right)^{k-1} du = \frac{t^{k+1}}{2\theta^k}$$

if $0 \le t \le \theta$. On the other hand, $P_\theta[T \le t] = (t/\theta)^k$, so that under P_θ the distribution of T has density kt^{k-1}/θ^k over $[0, \theta]$. Thus

$$(34.11) \qquad \int_{T \le t} \frac{k+1}{2k}T\, dP_\theta = \frac{k+1}{2k}\int_0^t uk\frac{u^{k-1}}{\theta^k}\, du = \frac{t^{k+1}}{2\theta^k}.$$

Since (34.10) and (34.11) agree, (34.9) follows by Theorem 34.1. ∎

The essential ideas in the proof of Theorem 34.6 are most easily understood through a preliminary consideration of special cases.

Lemma 1. *Suppose that* $[P_\theta: \theta \in \Theta]$ *is dominated by a probability measure P and that each P_θ has with respect to P a density g_θ that is measurable $\mathscr{G}$. Then $\mathscr{G}$ is sufficient, and $P[A\|\mathscr{G}]$ is a version of $P_\theta[A\|\mathscr{G}]$ for each θ in Θ.*

PROOF. For G in $\mathscr{G}$, (34.4) gives

$$\int_G P[A\|\mathscr{G}]\, dP_\theta = \int_G E[I_A\|\mathscr{G}]g_\theta\, dP = \int_G E[I_A g_\theta\|\mathscr{G}]\, dP$$

$$= \int_G I_A g_\theta\, dP = \int_{A \cap G} g_\theta\, dP = P_\theta(A \cap G).$$

Therefore, $P[A\|\mathscr{G}]$—the conditional probability calculated with respect to P—does serve as a version of $P_\theta[A\|\mathscr{G}]$ for each θ in Θ. Thus $\mathscr{G}$ is sufficient for the family

$[P_\theta: \theta \in \Theta]$—even for this family augmented by P (which might happen to lie in the family to start with). ∎

For the necessity, suppose first that the family is dominated by one of its members.

Lemma 2. *Suppose that* $[P_\theta: \theta \in \Theta]$ *is dominated by* P_{θ_0} *for some* $\theta_0 \in \Theta$. *If* $\mathscr{G}$ *is sufficient, then each* P_θ *has with respect to* P_{θ_0} *a density* g_θ *that is measurable* $\mathscr{G}$.

PROOF. Let $p(A, \omega)$ be the function in the definition of sufficiency, and take $P_\theta[A\|\mathscr{G}]_\omega = p(A, \omega)$ for all $A \in \mathscr{F}$, $\omega \in \Omega$, and $\theta \in \Theta$. Let d_θ be any density of P_θ with respect to P_{θ_0}. By a number of applications of (34.4),

$$\int_A E_{\theta_0}[d_\theta\|\mathscr{G}]\,dP_{\theta_0} = \int I_A E_{\theta_0}[d_\theta\|\mathscr{G}]\,dP_{\theta_0}$$

$$= \int E_{\theta_0}\{I_A E_{\theta_0}[d_\theta\|\mathscr{G}]\|\mathscr{G}\}\,dP_{\theta_0} = \int E_{\theta_0}\{I_A\|\mathscr{G}\}E_{\theta_0}[d_\theta\|\mathscr{G}]\,dP_{\theta_0}$$

$$= \int E_{\theta_0}[E_{\theta_0}\{I_A\|\mathscr{G}\}d_\theta\|\mathscr{G}]\,dP_{\theta_0} = \int E_{\theta_0}\{I_A\|\mathscr{G}\}d_\theta\,dP_{\theta_0}$$

$$= \int P_{\theta_0}[A\|\mathscr{G}]\,dP_\theta = \int P_\theta[A\|\mathscr{G}]\,dP_\theta = P_\theta(A),$$

the next-to-last equality by sufficiency (the integrand on either side being $p\,(A,\cdot\,)$). Thus $g_\theta = E_{\theta_0}[d_\theta\|\mathscr{G}]$, which is measurable $\mathscr{G}$, can serve as a density for P_θ with respect to P_{θ_0}. ∎

To complete the proof of Theorem 34.6 requires one more lemma of a technical sort.

Lemma 3. *If* $[P_\theta: \theta \in \Theta]$ *is dominated by a σ-finite measure, then it is equivalent to some finite or countably infinite subfamily.*

In many examples, the P_θ are all equivalent to each other, in which case the subfamily can be taken to consist of a single P_{θ_0}.

PROOF. Since μ is σ-finite, there is a finite or countable partition of Ω into $\mathscr{F}$-sets A_n such that $0 < \mu(A_n) < \infty$. Choose positive constants a_n, one for each A_n, in such a way that $\sum_n a_n < \infty$. The finite measure with value $\sum_n a_n \mu(A \cap A_n)/\mu(A_n)$ at A dominates μ. In proving the lemma it is therefore no restriction to assume the family dominated by a *finite* measure μ.

Each P_θ is dominated by μ and hence has a density f_θ with respect to it. Let $S_\theta = [\omega: f_\theta(\omega) > 0]$. Then $P_\theta(A) = P_\theta(A \cap S_\theta)$ for all A, and $P_\theta(A) = 0$ if and only if $\mu(A \cap S_\theta) = 0$. In particular, S_θ supports P_θ.

Call a set B in $\mathscr{F}$ a *kernel* if $B \subset S_\theta$ for some θ, and call a finite or countable union of kernels a *chain*. Let α be the supremum of $\mu(C)$ over chains C. Since μ is finite and a finite or countable union of chains is a chain, α is finite and $\mu(C) = \alpha$ for some chain C. Suppose that $C = \bigcup_n B_n$, where each B_n is a kernel, and suppose that $B_n \subset S_{\theta_n}$.

The problem is to show that $[P_\theta: \theta \in \Theta]$ is dominated by $[P_{\theta_n}: n = 1, 2, \dots]$ and hence equivalent to it. Suppose that $P_{\theta_n}(A) = 0$ for all n. Then $\mu(A \cap S_{\theta_n}) = 0$, as observed above. Since $C \subset \bigcup_n S_{\theta_n}$, $\mu(A \cap C) = 0$, and it follows that $P_\theta(A \cap C) = 0$

whatever θ may be. But suppose that $P_\theta(A - C) > 0$. Then $P_\theta((A - C) \cap S_\theta) = P_\theta(A - C)$ is positive, and so $(A - C) \cap S_\theta$ is a kernel, disjoint from C, of positive μ-measure; this is impossible because of the maximality of C. Thus $P_\theta(A - C)$ is 0 along with $P_\theta(A \cap C)$, and so $P_\theta(A) = 0$.   ∎

Suppose that $[P_\theta: \theta \in \Theta]$ is dominated by a σ-finite μ, as in Theorem 34.6, so that, by Lemma 3, it contains a finite or infinite sequence $P_{\theta_1}, P_{\theta_2}, \ldots$ equivalent to the entire family. Fix one such sequence, and choose positive constants c_n, one for each θ_n, in such a way that $\Sigma_n c_n = 1$. Now define a probability measure P on $\mathcal{F}$ by

$$(34.12) \qquad\qquad P(A) = \sum_n c_n P_{\theta_n}(A).$$

Clearly, P is equivalent to $[P_{\theta_1}, P_{\theta_2}, \ldots]$ and hence to $[P_\theta: \theta \in \Theta]$, and all three are dominated by μ.

$$(34.13) \qquad\qquad P = [P_{\theta_1}, P_{\theta_2}, \ldots] \equiv [P_\theta: \theta \in \Theta] \ll \mu.$$

Proof of Sufficiency in Theorem 34.6. If each P_θ has density $g_\theta h$ with respect to μ, then by the construction (34.12), P has density fh with respect to μ, where $f = \Sigma_n c_n g_{\theta_n}$. Put $r_\theta = g_\theta/f$ if $f > 0$, and $r_\theta = 0$ (say) if $f = 0$. If each g_θ is measurable $\mathcal{G}$, the same is true of f and hence of the r_θ. Since $P[f = 0] = 0$ and P is equivalent to the entire family, $P_\theta[f = 0] = 0$ for all θ. Therefore,

$$\int_A r_\theta \, dP = \int_A r_\theta fh \, d\mu = \int_{A \cap [f > 0]} r_\theta fh \, d\mu = \int_{A \cap [f > 0]} g_\theta h \, d\mu$$

$$= P_\theta(A \cap [f > 0]) = P_\theta(A).$$

Each P_θ thus has with respect to the probability measure P a density measurable $\mathcal{G}$, and it follows by Lemma 1 that $\mathcal{G}$ is sufficient   ∎

Proof of Necessity in Theorem 34.6. Let $p(A, \omega)$ be a function such that, for each A and θ, $p(A, \cdot)$ is a version of $P_\theta[A \| \mathcal{G}]$, as required by the definition of sufficiency. For P as in (34.12) and $G \in \mathcal{G}$,

$$(34.14) \qquad\qquad \int_G p(A, \omega) P(d\omega) = \sum_n c_n \int_G p(A, \omega) P_{\theta_n}(d\omega)$$

$$= \sum_n c_n \int_G P_{\theta_n}[A \| \mathcal{G}] \, dP_{\theta_n}$$

$$= \sum_n c_n P_{\theta_n}(A \cap G) = P(A \cap G).$$

Thus $p(A, \cdot)$ serves as a version of $P[A \| \mathcal{G}]$ as well, and $\mathcal{G}$ is still sufficient if P is added to the family. Since P dominates the augmented family, Lemma 2 implies that each P_θ has with respect to P a density g_θ that is measurable $\mathcal{G}$. But if h is the density of P with respect to μ (see (34.13)), then P_θ has density $g_\theta h$ with respect to μ.   ∎

A sub-σ-field $\mathscr{S}_0$ sufficient with respect to $[P_\theta: \theta \in \Theta]$ is *minimal* if, for each sufficient $\mathscr{S}$, $\mathscr{S}_0$ is essentially contained in $\mathscr{S}$ in the sense that for each A in $\mathscr{S}_0$ there is a B in $\mathscr{S}$ such that $P_\theta(A \vartriangle B) = 0$ for all θ in Θ. A sufficient $\mathscr{S}$ represents a compression of the information in $\mathscr{F}$, and a minimal sufficient $\mathscr{S}_0$ represents the greatest possible compression.

Suppose the densities f_θ of the P_θ with respect to μ have the property that $f_\theta(\omega)$ is measurable $\mathscr{E} \times \mathscr{F}$, where $\mathscr{E}$ is a σ-field in Θ. Let π be a probability measure on $\mathscr{E}$, and define P as $\int_\Theta P_\theta \pi(d\theta)$, in the sense that $P(A) = \int_\Theta \int_A f_\theta(\omega) \mu(d\omega) \pi(d\theta) = \int_\Theta P_\theta(A) \pi(d\theta)$. Obviously, $P \ll [P_\theta: \theta \in \Theta]$. Assume that

$$(34\ 15) \qquad [P_\theta: \theta \in \Theta] \ll P = \int_\Theta P_\theta \pi(d\theta).$$

If π has mass c_n at θ_n, then P is given by (34.12), and of course, (35.15) holds if (34.13) does. Let r_θ be a density for P_θ with respect to P.

Theorem 34.7. *If* (34.15) *holds, then* $\mathscr{S}_0 = \sigma[r_\theta: \theta \in \Theta]$ *is a minimal sufficient sub-σ-field.*

PROOF. That $\mathscr{S}_0$ is sufficient follows by Theorem 34.6. Suppose that $\mathscr{S}$ is sufficient. It follows by a simple extension of (34.14) that $\mathscr{S}$ is still sufficient if P is added to the family, and then it follows by Lemma 2 that each P_θ has with respect to P a density g_θ that is measurable $\mathscr{S}$. Since densities are essentially unique, $P[g_\theta = r_\theta] = 1$. Let $\mathscr{H}$ be the class of A in $\mathscr{S}_0$ such that $P(A \vartriangle B) = 0$ for some B in $\mathscr{S}$. Then $\mathscr{H}$ is a σ-field containing each set of the form $A = [r_\theta \in H]$ (take $B = [g_\theta \in H]$) and hence containing $\mathscr{S}_0$. Since, by (34.15), P dominates each P_θ, $\mathscr{S}_0$ is essentially contained in $\mathscr{S}$, in the sense of the definition. ∎

Minimum-Variance Estimation*

To illustrate sufficiency, let g be a real function on Θ, and consider the problem of estimating $g(\theta)$. One possibility is that Θ is a subset of the line and g is the identity; another is that Θ is a subset of R^k and g picks out one of the coordinates. (This problem is considered from a slightly different point of view at the end of Section 19.) An *estimate* of $g(\theta)$ is a random variable Z, and the estimate is *unbiased* if $E_\theta[Z] = g(\theta)$ for all θ. One measure of the accuracy of the estimate Z is $E_\theta[(Z - g(\theta))^2]$.

If $\mathscr{S}$ is sufficient, it follows by linearity (Theorem 34.2(ii)) that $E_\theta[X\|\mathscr{S}]$ has for simple X a version that is independent of θ. Since there are simple X_n such that $|X_n| \le |X|$ and $X_n \to X$, the same is true of any X that is integrable with respect to each P_θ (use Theorem 34.2(v)). Suppose that $\mathscr{S}$ is, in fact, sufficient, and denote by $E[X\|\mathscr{S}]$ a version of $E_\theta[X\|\mathscr{S}]$ that is independent of θ.

Theorem 34.8. *Suppose that* $E_\theta[(Z - g(\theta))^2] < \infty$ *for all* θ *and that* $\mathscr{S}$ *is sufficient. Then*

$$(34.16) \qquad E_\theta\left[(E[Z\|\mathscr{S}] - g(\theta))^2\right] \le E_\theta\left[(Z - g(\theta))^2\right]$$

for all θ. *If* Z *is unbiased, then so is* $E[Z\|\mathscr{S}]$.

---

*This topic may be omitted.

PROOF. By Jensens's inequality (34.7) for $\varphi(x) = (x - g(\theta))^2$, $(E[Z\|\mathscr{G}] - g(\theta))^2$
$\leq E_\theta[(Z - g(\theta))^2\|\mathscr{G}]$. Applying E_θ to each side gives (34.16). The second statement
follows from the fact that $E_\theta[E[Z\|\mathscr{G}]] = E_\theta[Z]$.   ∎

This, the *Rao–Blackwell theorem*, says that $E[Z\|\mathscr{G}]$ is at least as good an estimate
as Z if $\mathscr{G}$ is sufficient.

Example 34.5. Returning to Example 34.4, note that each X_i has mean $\theta/2$
under P_θ, so that if $\bar{X} = k^{-1}\sum_{i=1}^{k}X_i$ is the sample mean, then $2\bar{X}$ is an unbiased
estimate of θ. But there is a better one. By (34.9), $E_\theta[2\bar{X}\|T] = (k + 1)T/k = T'$, and
by the Rao–Blackwell theorem, T' is an unbiased estimate with variance at most that
of $2\bar{X}$.

In fact, for an arbitrary unbiased estimate Z, $E_\theta[(T' - \theta)^2] \leq E_\theta[(Z - \theta)^2]$. To
prove this, let $\delta = T' - E[Z\|T]$. By Theorem 20.1(ii), $\delta = f(T)$ for some Borel
function f, and $E_\theta[f(T)] = 0$ for all θ. Taking account of the density for T leads to
$\int_0^\theta f(x)x^{k-1}\,dx = 0$, so that $f(x)x^{k-1}$ integrates to 0 over all intervals. Therefore,
$f(x)$ along with $f(x)x^{k-1}$ vanishes for $x > 0$, except on a set of Lebesgue measure 0,
and hence $P_\theta[f(T) = 0] = 1$ and $P_\theta[T' = E[Z\|T]] = 1$ for all θ. Therefore, $E_\theta[(T' - \theta)^2] = E_\theta[(E[Z\|T] - \theta)^2] \leq E_\theta[(Z - \theta)^2]$ for Z unbiased, and T' has minimum variance among all unbiased estimates of θ.   ∎

PROBLEMS

34.1. Work out for conditional expected values the analogues of Problems 33.4, 33.5,
and 33.9.

34.2. In the context of Examples 33.5 and 33.12, show that the conditional expected
value of Y (if it is integrable) given X is $g(X)$, where

$$g(x) = \frac{\int_{-\infty}^{\infty} f(x, y)y\,dy}{\int_{-\infty}^{\infty} f(x, y)\,dy}.$$

34.3. Show that the independence of X and Y implies that $E[Y\|X] = E[Y]$, which
in turn implies that $E[XY] = E[X]E[Y]$. Show by examples in an Ω of three
points that the reverse implications are both false.

34.4. (a) Let B be an event with $P(B) > 0$, and define a probability measure P_0 by
$P_0(A) = P(A|B)$. Show that $P_0[A\|\mathscr{G}] = P[A \cap B\|\mathscr{G}]/P[B\|\mathscr{G}]$ on a set of
P_0-measure 1.

(b) Suppose that $\mathscr{H}$ is generated by a partition $B_1, B_2, \ldots$, and let $\mathscr{G} \vee \mathscr{H} = \sigma(\mathscr{G} \cup \mathscr{H})$. Show that with probability 1,

$$P[A\|\mathscr{G} \vee \mathscr{H}] = \sum_i I_{B_i} \frac{P[A \cap B_i\|\mathscr{G}]}{P[B_i\|\mathscr{G}]}.$$

34.5. The equation (34.5) was proved by showing that the left side is a version of the right side. Prove it by showing that the right side is a version of the left side.

34.6. Prove for bounded X and Y that $E[YE[X\|\mathscr{G}]] = E[XE[Y\|\mathscr{G}]]$.

34.7. 33.9↑   Generalize Theorem 34.5 by replacing X with a random vector

34.8. Assume that X is nonnegative but not necessarily integrable. Show that it is still possible to define a nonnegative random variable $E[X\|\mathscr{G}]$, measurable $\mathscr{G}$, such that (34.1) holds. Prove versions of the monotone convergence theorem and Fatou's lemma.

34.9. (a) Show for nonnegative X that $E[X\|\mathscr{G}] = \int_0^\infty P[X > t\|\mathscr{G}]\,dt$ with probability 1.

   (b) Generalize Markov's inequality: $P\{|X| \geq \alpha\|\mathscr{G}\} \leq \alpha^{-k}E[|X|^k\|\mathscr{G}]$ with probability 1.

   (c) Similarly generalize Chevyshev's and Hölder's inequalities.

34.10. (a) Show that, if $\mathscr{G}_1 \subset \mathscr{G}_2$ and $E[X^2] < \infty$, then $E[(X - E[X\|\mathscr{G}_2])^2] \leq E[(X - E[X\|\mathscr{G}_1])^2]$. The dispersion of X about its conditional mean becomes smaller as the σ-field grows.

   (b) Define $\mathrm{Var}[X\|\mathscr{G}] = E[(X - E[X\|\mathscr{G}])^2\|\mathscr{G}]$. Prove that $\mathrm{Var}[X] = E[\mathrm{Var}[X\|\mathscr{G}]] + \mathrm{Var}[E[X\|\mathscr{G}]]$.

34.11. Let $\mathscr{G}_1, \mathscr{G}_2, \mathscr{G}_3$ be σ-fields in $\mathscr{F}$, let $\mathscr{G}_{ij}$ be the σ-field generated by $\mathscr{G}_i \cup \mathscr{G}_j$, and let A_i be the generic set in $\mathscr{G}_i$. Consider three conditions:
   (i) $P[A_3\|\mathscr{G}_{12}] = P[A_3\|\mathscr{G}_2]$ for all A_3.
   (ii) $P[A_1 \cap A_3\|\mathscr{G}_2] = P[A_1\|\mathscr{G}_2]P[A_3\|\mathscr{G}_2]$ for all A_1 and A_3.
   (iii) $P[A_1\|\mathscr{G}_{23}] = P[A_1\|\mathscr{G}_2]$ for all A_1.
   If $\mathscr{G}_1$, $\mathscr{G}_2$, and $\mathscr{G}_3$ are interpreted as descriptions of the past, present, and future, respectively, (i) is a general version of the Markov property: the conditional probability of a future event A_3 given the past and present $\mathscr{G}_{12}$ is the same as the conditional probability given the present $\mathscr{G}_2$ alone. Condition (iii) is the same with time reversed. And (ii) says that past and future events A_1 and A_3 are conditionally independent given the present $\mathscr{G}_2$. Prove the three conditions equivalent.

34.12. 33.7   34.11↑   Use Example 33.10 to calculate $P[N_s = k\|N_u, u \geq t]$ $(s \leq t)$ for the Poisson process.

34.13. Let L^2 be the Hilbert space of square-integrable random variables on $(\Omega, \mathscr{F}, P)$. For $\mathscr{G}$ a σ-field in $\mathscr{F}$, let $M_\mathscr{G}$ be the subspace of elements of L^2 that are measurable $\mathscr{G}$. Show that the operator $P_\mathscr{G}$ defined for $X \in L^2$ by $P_\mathscr{G}X = E[X\|\mathscr{G}]$ is the perpendicular projection on $M_\mathscr{G}$.

34.14. ↑   Suppose in Problem 34.13 that $\mathscr{G} = \sigma(Z)$ for a random variable Z in L^2. Let S_Z be the one-dimensional subspace spanned by Z. Show that S_Z may be much smaller than $M_{\sigma(Z)}$, so that $E[X\|Z]$ (for $X \in L^2$) is by no means the projection of X on Z. *Hint*: Take Z the identity function on the unit interval with Lebesgue measure.

34.15. ↑ Problem 34.13 can be turned around to give an alternative approach to conditional probability and expected value. For a σ-field $\mathscr{G}$ in $\mathscr{F}$, let $P_{\mathscr{G}}$ be the perpendicular projection on the subspace $M_{\mathscr{G}}$. Show that $P_{\mathscr{G}}X$ has for $X \in L^2$ the two properties required of $E[X\|\mathscr{G}]$. Use this to *define* $E[X\|\mathscr{G}]$ for $X \in L^2$ and then extend it to all integrable X via approximation by random variables in L^2. Now define conditional probability.

34.16. *Mixing sequences.* A sequence $A_1, A_2,\ldots$ of $\mathscr{F}$-sets in a probability space $(\Omega, \mathscr{F}, P)$ is *mixing* with constant α if

(34.17) $$\lim_n P(A_n \cap E) = \alpha P(E)$$

for every E in $\mathscr{F}$. Then $\alpha = \lim_n P(A_n)$.

(a) Show that $\{A_n\}$ is mixing with constant α if and only if

(34.18) $$\lim_n \int_{A_n} X\,dP = \alpha \int X\,dP$$

for each integrable X (measurable $\mathscr{F}$).

(b) Suppose that (34.17) holds for $E \in \mathscr{P}$, where $\mathscr{P}$ is a π-system, $\Omega \in \mathscr{P}$, and $A_n \in \sigma(\mathscr{P})$ for all n. Show that $\{A_n\}$ is mixing. *Hint*: First check (34.18) for X measurable $\sigma(\mathscr{P})$ and then use conditional expected values with respect to $\sigma(\mathscr{P})$.

(c) Show that, if P_0 is a probability measure on $(\Omega, \mathscr{F})$ and $P_0 \ll P$, then mixing is preserved if P is replaced by P_0.

34.17. ↑ *Application of mixing to the central limit theorem.* Let $X_1, X_2,\ldots$ be random variables on $(\Omega, \mathscr{F}, P)$, independent and identically distributed with mean 0 and variance σ^2, and put $S_n = X_1 + \cdots + X_n$. Then $S_n/\sigma\sqrt{n} \Rightarrow N$ by the Lindeberg-Lévy theorem. Show by the steps below that this still holds if P is replaced by any probability measure P_0 on $(\Omega, \mathscr{F})$ that P dominates. For example, the central limit theorem applies to the sums $\sum_{k=1}^n r_k(\omega)$ of Rademacher functions if ω is chosen according to the uniform density over the unit interval, and this result shows that the same is true if ω is chosen according to an arbitrary density.

Let $Y_n = S_n/\sigma\sqrt{n}$ and $Z_n = (S_n - S_{[\log n]})/\sigma\sqrt{n}$, and take $\mathscr{P}$ to consist of the sets of the form $[(X_1,\ldots,X_k) \in H]$, $k \geq 1$, $H \in \mathscr{R}^k$. Prove successively:

(a) $P[Y_n \leq x] \to P[N \leq x]$.

(b) $P[|Y_n - Z_n| \geq \epsilon] \to 0$.

(c) $P[Z_n \leq x] \to P[N \leq x]$.

(d) $P(E \cap [Z_n \leq x]) \to P(E)P[N \leq x]$ for $E \in \mathscr{P}$.

(e) $P(E \cap [Z_n \leq x]) \to P(E)P[N \leq x]$ for $E \in \mathscr{F}$.

(f) $P_0[Z_n \leq x] \to P[N \leq x]$.

(g) $P_0[|Y_n - Z_n| \geq \epsilon] \to 0$.

(h) $P_0[Y_n \leq x] \to P[N \leq x]$.

34.18. Suppose that $\mathscr{G}$ is a sufficient subfield for the family of probability measures P_θ, $\theta \in \Theta$, on $(\Omega, \mathscr{F})$. Suppose that for each θ and A, $p(A, \omega)$ is a version of $P_\theta[A\|\mathscr{G}]_\omega$, and suppose further that for each ω, $p(\cdot, \omega)$ is a probability

measure on $\mathscr{F}$. Define Q_θ on $\mathscr{F}$ by $Q_\theta(A) = \int_\Omega p(A, \omega)P_\theta(d\omega)$, and show that $Q_\theta = P_\theta$.

The idea is that an observer with the information in $\mathscr{G}$ (but ignorant of ω itself) in principle knows the values $p(A, \omega)$ because each $p(A, \cdot)$ is measurable $\mathscr{G}$. If he has the appropriate randomization device, he can draw an ω' from Ω according to the probability measure $p(\cdot, \omega)$, and his ω' will have the same distribution $Q_\theta = P_\theta$ that ω has. Thus, whatever the value of the unknown θ, the observer can on the basis of the information in $\mathscr{G}$ alone, and without knowing ω itself, construct a probabilistic replica of ω.

34.19. 34.13 ↑   In the context of the discussion on p. 252, let $\bar{\mathscr{F}}$ be the σ-field of sets of the form $\Theta \times A$ for $A \in \mathscr{F}$. Show that under the probability measure Q, t_0 is the conditional expected value of $\bar{g}_0$ given $\bar{\mathscr{F}}$.

34.20. (a) In Example 34.4, take π to have density $e^{-\theta}$ over $\Theta = (0, \infty)$. Show by Theorem 34.7 that T is a minimal sufficient statistic (in the sense that $\sigma(T)$ is minimal).

(b) Let P_θ be the distribution for samples of size n from a normal distribution with parameter $\theta = (m, \sigma^2)$, $\sigma^2 > 0$, and let π put unit mass at $(0, 1)$. Show that the sample mean and variance form a minimal sufficient statistic.

SECTION 35. MARTINGALES

Definition

Let $X_1, X_2, \ldots$ be a sequence of random variables on a probability space $(\Omega, \mathscr{F}, P)$, and let $\mathscr{F}_1, \mathscr{F}_2, \ldots$ be a sequence of σ-fields in $\mathscr{F}$. The sequence $\{(X_n, \mathscr{F}_n): n = 1, 2, \ldots\}$ is a *martingale* if these four conditions hold:

(i) $\mathscr{F}_n \subset \mathscr{F}_{n+1}$;
(ii) X_n *is measurable* $\mathscr{F}_n$;
(iii) $E[|X_n|] < \infty$;
(iv) *with probability* 1,

$$(35.1) \qquad\qquad E[X_{n+1}\|\mathscr{F}_n] = X_n.$$

Alternatively, the sequence $X_1. X_2, \ldots$ is said to be a *martingale relative to* the σ-fields $\mathscr{F}_1, \mathscr{F}_2, \ldots$. Condition (i) is expressed by saying the $\mathscr{F}_n$ form a *filtration* and condition (ii) by saying the X_n are *adapted* to the filtration.

If X_n represents the fortune of a gambler after the nth play and $\mathscr{F}_n$ represents his information about the game at that time, (35.1) says that his expected fortune after the next play is the same as his present fortune. Thus a martingale represents a fair game, and sums of independent random variables with mean 0 give one example. As will be seen below, martingales arise in very diverse connections.

The sequence $X_1, X_2, \ldots$ is defined to be a martingale if it is a martingale relative to *some* sequence $\mathscr{F}_1, \mathscr{F}_2, \ldots$. In this case, the σ-fields $\mathscr{G}_n = \sigma(X_1, \ldots, X_n)$ always work: Obviously, $\mathscr{G}_n \subset \mathscr{G}_{n+1}$ and X_n is measurable $\mathscr{G}_n$, and if (35.1) holds, then $E[X_{n+1} \| \mathscr{G}_n] = E[E[X_{n+1} \| \mathscr{F}_n] \| \mathscr{G}_n] = E[X_n \| \mathscr{G}_n] = X_n$ by (34.5). For these special σ-fields $\mathscr{G}_n$, (35.1) reduces to

$$(35.2) \qquad E[X_{n+1} \| X_1, \ldots, X_n] = X_n.$$

Since $\sigma(X_1, \ldots, X_n) \subset \mathscr{F}_n$ if and only if X_n is measurable $\mathscr{F}_n$ for each n, the $\sigma(X_1, \ldots, X_n)$ are the *smallest* σ-fields with respect to which the X_n are a martingale.

The essential condition is embodied in (35.1) and in its specialization (35.2). Condition (iii) is of course needed to ensure that $E[X_{n+1} \| \mathscr{F}_n]$ exists. Condition (iv) says that X_n is a version of $E[X_{n+1} \| \mathscr{F}_n]$; since X_n is measurable $\mathscr{F}_n$, the requirement reduces to

$$(35.3) \qquad \int_A X_{n+1} \, dP = \int_A X_n \, dP, \qquad A \in \mathscr{F}_n.$$

Since the $\mathscr{F}_n$ are nested, $A \in \mathscr{F}_n$ implies that $\int_A X_n \, dP = \int_A X_{n+1} \, dP = \cdots = \int_A X_{n+k} \, dP$. Therefore, X_n, being measurable $\mathscr{F}_n$, is a version of $E[X_{n+k} \| \mathscr{F}_n]$:

$$(35.4) \qquad E[X_{n+k} \| \mathscr{F}_n] = X_n$$

with probability 1 for $k \geq 1$. Note that for $A = \Omega$, (35.3) gives

$$(35.5) \qquad E[X_1] = E[X_2] = \cdots.$$

The defining conditions for a martingale can also be given in terms of the differences

$$(35.6) \qquad \Delta_n = X_n - X_{n-1}$$

$(\Delta_1 = X_1)$. By linearity, (35.1) is the same thing as

$$(35.7) \qquad E[\Delta_{n+1} \| \mathscr{F}_n] = 0.$$

Note that, since $X_k = \Delta_1 + \cdots + \Delta_k$ and $\Delta_k = X_k - X_{k-1}$, the sets $X_1, \ldots, X_n$ and $\Delta_1, \ldots, \Delta_n$ generate the same σ-field:

$$(35.8) \qquad \sigma(X_1, \ldots, X_n) = \sigma(\Delta_1, \ldots, \Delta_n).$$

Example 35.1. Let $\Delta_1, \Delta_2, \ldots$ be independent, integrable random variables such that $E[\Delta_n] = 0$ for $n \geq 2$. If $\mathscr{F}_n$ is the σ-field (35.8), then by independence $E[\Delta_{n+1} \| \mathscr{F}_n] = E[\Delta_{n+1}] = 0$. If Δ is another random variable,

independent of the Δ_n, and if $\mathscr{F}_n$ is replaced by $\sigma(\Delta, \Delta_1, \ldots, \Delta_n)$, then the $X_n = \Delta_1 + \cdots + \Delta_n$ are still a martingale relative to the $\mathscr{F}_n$. It is natural and convenient in the theory to allow σ-fields $\mathscr{F}_n$ larger than the minimal ones (35.8). ∎

Example 35.2. Let $(\Omega, \mathscr{F}, P)$ be a probability space, let ν be a finite measure on $\mathscr{F}$, and let $\mathscr{F}_1, \mathscr{F}_2, \ldots$ be a nondecreasing sequence of σ-fields in $\mathscr{F}$. Suppose that P dominates ν when both are restricted to $\mathscr{F}_n$—that is, suppose that $A \in \mathscr{F}_n$ and $P(A) = 0$ together imply that $\nu(A) = 0$. There is then a density or Radon–Nikodym derivative X_n of ν with respect to P when both are restricted to $\mathscr{F}_n$; X_n is a function that is measurable $\mathscr{F}_n$ and integrable with respect to P, and it satisfies

$$(35.9) \qquad \int_A X_n \, dP = \nu(A), \qquad A \in \mathscr{F}_n.$$

If $A \in \mathscr{F}_n$, then $A \in \mathscr{F}_{n+1}$ as well, so that $\int_A X_{n+1} \, dP = \nu(A)$; this and (35.9) give (35.3). Thus the X_n are a martingale with respect to the $\mathscr{F}_n$. ∎

Example 35.3. For a specialization of the preceding example, let P be Lebesgue measure on the σ-field $\mathscr{F}$ of Borel subsets of $\Omega = (0, 1]$, and let $\mathscr{F}_n$ be the finite σ-field generated by the partition of Ω into dyadic intervals $(k2^{-n}, (k+1)2^{-n}]$, $0 \le k < 2^n$. If $A \in \mathscr{F}_n$ and $P(A) = 0$, then A is empty. Hence P dominates *every* finite measure ν on $\mathscr{F}_n$. The Radon–Nikodym derivative is

$$(35.10) \quad X_n(\omega) = \frac{\nu(k2^{-n}, (k+1)2^{-n}]}{2^{-n}} \qquad \text{if } \omega \in (k2^{-n}, (k+1)2^{-n}].$$

There is no need here to assume that P dominates ν when they are viewed as measures on all of $\mathscr{F}$. Suppose that ν is the distribution of $\sum_{k=1}^{\infty} Z_k 2^{-k}$ for independent Z_k assuming values 1 and 0 with probabilities p and $1 - p$. This is the measure in Examples 31.1 and 31.3 (there denoted by μ), and for $p \ne \frac{1}{2}$, ν is singular with respect to Lebesgue measure P. It is nonetheless absolutely continuous with respect to P when both are restricted to $\mathscr{F}_n$. ∎

Example 35.4. For another specialization of Example 35.2, suppose that ν is a probability measure Q on $\mathscr{F}$ and that $\mathscr{F}_n = \sigma(Y_1, \ldots, Y_n)$ for random variables $Y_1, Y_2, \ldots$ on $(\Omega, \mathscr{F})$. Suppose that under the measure P the distribution of the random vector $(Y_1, \ldots, Y_n)$ has density $p_n(y_1, \ldots, y_n)$ with respect to n-dimensional Lebesgue measure and that under Q it has density $q_n(y_1, \ldots, y_n)$. To avoid technicalities, assume that p_n is everywhere positive.

Then the Radon–Nikodym derivative for Q with respect to P on $\mathscr{F}_n$ is

$$(35.11) \qquad X_n = \frac{q_n(Y_1, \ldots, Y_n)}{p_n(Y_1, \ldots, Y_n)}.$$

To see this, note that the general element of $\mathscr{F}_n$ is $[(Y_1, \ldots, Y_n) \in H]$, $H \in \mathscr{R}^n$; by the change-of-variable formula,

$$\int_{[(Y_1, \ldots, Y_n) \in H]} X_n \, dP = \int_H \frac{q_n(y_1, \ldots, y_n)}{p_n(y_1, \ldots, y_n)} p_n(y_1, \ldots, y_n) \, dy_1 \cdots dy_n$$

$$= Q[(Y_1, \ldots, Y_n) \in H].$$

In statistical terms, (35.11) is a likelihood ratio: p_n and q_n are rival densities, and the larger X_n is, the more strongly one prefers q_n as an explanation of the observation $(Y_1, \ldots, Y_n)$. The analysis is carried out under the assumption that P is the measure actually governing the Y_n; that is, X_n is a martingale under P and not in general under Q.

In the most common case the Y_n are independent and identically distributed under both P and Q: $p_n(y_1, \ldots, y_n) = p(y_1) \cdots p(y_n)$ and $q_n(y_1, \ldots, y_n) = q(y_1) \cdots q(y_n)$ for densities p and q on the line, where p is assumed everywhere positive for simplicity. Suppose that the measures corresponding to the densities p and q are not identical, so that $P[Y_n \in H] \ne Q[Y_n \in H]$ for some $H \in \mathscr{R}^1$. If $Z_n = I_{[Y_n \in H]}$, then by the strong law of large numbers, $n^{-1} \sum_{k=1}^n Z_k$ converges to $P[Y_1 \in H]$ on a set (in $\mathscr{F}$) of P-measure 1 and to $Q[Y_1 \in H]$ on a (disjoint) set of Q-measure 1. Thus P and Q are mutually singular on $\mathscr{F}$ even though P dominates Q on $\mathscr{F}_n$. ∎

Example 35.5. Suppose that Z is an integrable random variable on $(\Omega, \mathscr{F}, P)$ and that $\mathscr{F}_n$ are nondecreasing σ-fields in $\mathscr{F}$. If

$$(35.12) \qquad X_n = E[Z \| \mathscr{F}_n],$$

then the first three conditions in the martingale definition are satisfied, and by (34.5), $E[X_{n+1} \| \mathscr{F}_n] = E[E[Z \| \mathscr{F}_{n+1}] \| \mathscr{F}_n] = E[Z \| \mathscr{F}_n] = X_n$. Thus X_n is a martingale relative to $\mathscr{F}_n$. ∎

Example 35.6. Let N_{nk}, $n, k = 1, 2, \ldots$, be an independent array of identically distributed random variables assuming the values $0, 1, 2, \ldots$. Define $Z_0, Z_1, Z_2, \ldots$ inductively by $Z_0(\omega) = 1$ and $Z_n(\omega) = N_{n,1}(\omega) + \cdots + N_{n, Z_{n-1}(\omega)}(\omega)$; $Z_n(\omega) = 0$ if $Z_{n-1}(\omega) = 0$. If N_{nk} is thought of as the number of progeny of an organism, and if Z_{n-1} represents the size at time $n-1$ of a population of these organisms, then Z_n represents the size at time n. If the expected number of progeny is $E[N_{nk}] = m$, then $E[Z_n \| Z_{n-1}] = Z_{n-1}m$, so that $X_n = Z_n/m^n$, $n = 0, 1, 2, \ldots$, is a martingale. The sequence $Z_0, Z_1, \ldots$ is a *branching process*. ∎

In the preceding definition and examples, n ranges over the positive integers. The definition makes sense if n ranges over $1, 2, \ldots, N$; here conditions (ii) and (iii) are required for $1 \le n \le N$ and conditions (i) and (iv) only for $1 \le n < N$. It is, in fact, clear that the definition makes sense if the indices range over an arbitrary ordered set. Although martingale theory with an interval of the line as the index set is of great interest and importance, here the index set will be discrete.

Submartingales

Random variables X_n are a *submartingale* relative to σ-fields $\mathscr{F}_n$ if (i), (ii), and (iii) of the definition above hold and if this condition holds in place of (iv).

(iv') *with probability* 1,

$$(35.13) \qquad\qquad E[X_{n+1}\|\mathscr{F}_n] \ge X_n.$$

As before, the X_n are a submartingale if they are a submartingale with respect to some sequence $\mathscr{F}_n$, and the special sequence $\mathscr{F}_n = \sigma(X_1, \ldots, X_n)$ works if any does. The requirement (35.13) is the same thing as

$$(35.14) \qquad\qquad \int_A X_{n+1}\, dP \ge \int_A X_n\, dP, \qquad A \in \mathscr{F}_n.$$

This extends inductively (see the argument for (35.4)), and so

$$(35.15) \qquad\qquad E[X_{n+k}\|\mathscr{F}_n] \ge X_n$$

for $k \ge 1$. Taking expected values in (35.15) gives

$$(35.16) \qquad\qquad E[X_1] \le E[X_2] \le \cdots.$$

Example 35.7. Suppose that the Δ_n are independent and integrable, as in Example 35.1, but assume that $E[\Delta_n]$ is for $n \ge 2$ nonnegative rather than 0. Then the partial sums $\Delta_1 + \cdots + \Delta_n$ form a submartingale. ∎

Example 35.8. Suppose that the X_n are a martingale relative to the $\mathscr{F}_n$. Then $|X_n|$ is measurable $\mathscr{F}_n$ and integrable, and by Theorem 34.2(iv), $E[|X_{n+1}|\|\mathscr{F}_n] \ge |E[X_{n+1}\|\mathscr{F}_n]| = |X_n|$. Thus the $|X_n|$ are a submartingale relative to the $\mathscr{F}_n$. Note that even if $X_1, \ldots, X_n$ generate $\mathscr{F}_n$, in general $|X_1|, \ldots, |X_n|$ will generate a σ-field smaller than $\mathscr{F}_n$. ∎

Reversing the inequality in (35.13) gives the definition of a *supermartingale*. The inequalities in (35.14), (35.15), and (35.16) become reversed as well. The theory for supermartingales is of course symmetric to that of submartingales.

Gambling

Consider again the gambler whose fortune after the nth play is X_n and whose information about the game at that time is represented by the σ-field $\mathscr{F}_n$. If $\mathscr{F}_n = \sigma(X_1, \ldots, X_n)$, he knows the sequence of his fortunes and nothing else, but $\mathscr{F}_n$ could be larger. The martingale condition (35.1) stipulates that his expected or average fortune after the next play equals his present fortune, and so the martingale is the model for a *fair game*. Since the condition (35.13) for a submartingale stipulates that he stands to gain (or at least not lose) on the average, a submartingale represents a game *favorable* to the gambler. Similarly, a supermartingale represents a game *unfavorable* to the gambler.[†]

Examples of such games were studied in Section 7, and some of the results there have immediate generalizations. Start the martingale at $n = 0$, X_0 representing the gambler's initial fortune. The difference $\Delta_n = X_n - X_{n-1}$ represents the amount the gambler wins on the nth play,[‡] a negative win being of course a loss. Suppose instead that Δ_n represents the amount he wins if he puts up unit stakes. If instead of unit stakes he wagers the amount W_n on the nth play, $W_n \Delta_n$ represents his gain on that play. Suppose that $W_n \geq 0$, and that W_n is measurable $\mathscr{F}_{n-1}$ to exclude prevision: Before the nth play the information available to the gambler is that in $\mathscr{F}_{n-1}$, and his choice of stake W_n must be based on this alone. For simplicity take W_n bounded. Then $W_n \Delta_n$ is integrable, and it is measurable $\mathscr{F}_n$ if Δ_n is, and if X_n is a martingale, then $E[W_n \Delta_n \| \mathscr{F}_{n-1}] = W_n E[\Delta_n \| \mathscr{F}_{n-1}] = 0$ by (34.2). Thus

$$(35.17) \qquad X_0 + W_1 \Delta_1 + \cdots + W_n \Delta_n$$

is a martingale relative to the $\mathscr{F}_n$. The sequence $W_1, W_2, \ldots$ represents a betting system, and transforming a fair game by a betting system preserves fairness; that is, transforming X_n into (35.17) preserves the martingale property.

The various betting systems discussed in Section 7 give rise to various martingales, and these martingales are not in general sums of independent random variables—are not in general the special martingales of Example 35.1. If W_n assumes only the values 0 and 1, the betting system is a selection system; see Section 7.

If the game is unfavorable to the gambler—that is, if X_n is a supermartingale—and if W_n is nonnegative, bounded, and measurable $\mathscr{F}_{n-1}$, then the same argument shows that (35.17) is again a supermartingale, is again unfavorable. Betting systems are thus of no avail in unfavorable games.

The stopping-time arguments of Section 7 also extend. Suppose that $\{X_n\}$ is a martingale relative to $\{\mathscr{F}_n\}$; it may have come from another martingale

---

[†]There is a reversal of terminology here: a subfair game (Section 7) is against the gambler, while a submartingale favors him.

[‡]The notation has, of course, changed. The F_n and X_n of Section 7 have become X_n and Δ_n.

via transformation by a betting system. Let τ be a random variable taking on nonnegative integers as values, and suppose that

(35.18) $[\tau = n] \in \mathscr{F}_n.$

If τ is the time the gambler stops, $[\tau = n]$ is the event he stops just after the nth play, and (35.18) requires that his decision is to depend only on the information $\mathscr{F}_n$ available to him at that time. His fortune at time n for this stopping rule is

(35.19) $X_n^* = \begin{cases} X_n & \text{if } n \leq \tau, \\ X_\tau & \text{if } n \geq \tau. \end{cases}$

Here X_τ (which has value $X_{\tau(\omega)}(\omega)$ at ω) is the gambler's ultimate fortune, and it is his fortune for all times subsequent to τ.

The problem is to show that $X_0^*, X_1^*, \ldots$ is a martingale relative to $\mathscr{F}_0, \mathscr{F}_1, \ldots$. First,

$$E[|X_n^*|] = \sum_{k=0}^{n-1} \int_{[\tau=k]} |X_k|\, dP + \int_{[\tau \geq n]} |X_n|\, dP \leq \sum_{k=0}^{n} E[|X_k|] < \infty.$$

Since $[\tau > n] = \Omega - [\tau \leq n] \in \mathscr{F}_n,$

$$[X_n^* \in H] = \bigcup_{k=0}^{n} [\tau = k,\, X_k \in H] \cup [\tau > n,\, X_n \in H] \in \mathscr{F}_n.$$

Moreover,

$$\int_A X_n^* \, dP = \int_{A \cap [\tau > n]} X_n \, dP + \int_{A \cap [\tau \leq n]} X_\tau \, dP$$

and

$$\int_A X_{n+1}^* \, dP = \int_{A \cap [\tau > n]} X_{n+1} \, dP + \int_{A \cap [\tau \leq n]} X_\tau \, dP.$$

Because of (35.3), the right sides here coincide if $A \in \mathscr{F}_n$; this establishes (35.3) for the sequence $X_1^*, X_2^*, \ldots$, which is thus a martingale. The same kind of argument works for supermartingales.

Since $X_n^* = X_\tau$ for $n \geq \tau$, $X_n^* \to X_\tau$. As pointed out in Section 7, it is not always possible to integrate to the limit here. Let $X_n = a + \Delta_1 + \cdots + \Delta_n$, where the Δ_n are independent and assume the values ± 1 with probability $\frac{1}{2}$ ($X_0 = a$), and let τ be the smallest n for which $\Delta_1 + \cdots + \Delta_n = 1$. Then $E[X_0^*] = a$ and $X_\tau = a + 1$. On the other hand, if the X_n are uniformly bounded or uniformly integrable, it is possible to integrate to the limit: $E[X_\tau] = E[X_0]$.

Functions of Martingales

Convex functions of martingales are submartingales:

Theorem 35.1. (i) *If $X_1, X_2, \ldots$ is a martingale relative to $\mathscr{F}_1, \mathscr{F}_2, \ldots$, if φ is convex, and if the $\varphi(X_n)$ are integrable, then $\varphi(X_1), \varphi(X_2), \ldots$ is a submartingale relative to $\mathscr{F}_1, \mathscr{F}_2$.*

(ii) *If $X_1, X_2, \ldots$ is a submartingale relative to $\mathscr{F}_1, \mathscr{F}_2, \ldots$, if φ is nondecreasing and convex, and if the $\varphi(X_n)$ are integrable, then $\varphi(X_1), \varphi(X_2), \ldots$ is a submartingale relative to $\mathscr{F}_1, \mathscr{F}_2, \ldots$.*

PROOF. In the submartingale case, $X_n \leq E[X_{n+1} \| \mathscr{F}_n]$, and if φ is nondecreasing, then $\varphi(X_n) \leq \varphi(E[X_{n+1} \| \mathscr{F}_n])$. In the martingale case, $X_n = E[X_{n+1} \| \mathscr{F}_n]$, and so $\varphi(X_n) = \varphi(E[X_{n+1} \| \mathscr{F}_n])$. If φ is convex, then by Jensen's inequality (34.7) for conditional expectations, it follows that $\varphi(E[X_{n+1} \| \mathscr{F}_n]) \leq E[\varphi(X_{n+1}) \| \mathscr{F}_n]$. ∎

Example 35.8 is the case of part (i) for $\varphi(x) = |x|$.

Stopping Times

Let τ be a random variable taking as values positive integers or the special value ∞. It is a *stopping time* with respect to $\{\mathscr{F}_n\}$ if $[\tau = k] \in \mathscr{F}_k$ for each finite k (see (35.18)), or, equivalently, if $[\tau \leq k] \in \mathscr{F}_k$ for each finite k. Define

$$(35.20) \qquad \mathscr{F}_\tau = [A \in \mathscr{F} \colon A \cap [\tau \leq k] \in \mathscr{F}_k, 1 \leq k < \infty].$$

This is a σ-field, and the definition is unchanged if $[\tau \leq k]$ is replaced by $[\tau = k]$ on the right. Since clearly $[\tau = j] \in \mathscr{F}_\tau$ for finite j, τ is measurable $\mathscr{F}_\tau$.

If $\tau(\omega) < \infty$ for all ω and $\mathscr{F}_n = \sigma(X_1, \ldots, X_n)$, then $I_A(\omega) = I_A(\omega')$ for all A in $\mathscr{F}_\tau$ if and only if $X_i(\omega) = X_i(\omega')$ for $i \leq \tau(\omega) = \tau(\omega')$: The information in $\mathscr{F}_\tau$ consists of the values $\tau(\omega), X_1(\omega), \ldots, X_{\tau(\omega)}(\omega)$.

Suppose now that τ_1 and τ_2 are two stopping times and $\tau_1 \leq \tau_2$. If $A \in \mathscr{F}_{\tau_1}$, then $A \cap [\tau_1 \leq k] \in \mathscr{F}_k$ and hence $A \cap [\tau_2 \leq k] = A \cap [\tau_1 \leq k] \cap [\tau_2 \leq k] \in \mathscr{F}_k$: $\mathscr{F}_{\tau_1} \subset \mathscr{F}_{\tau_2}$.

Theorem 35.2. *If $X_1, \ldots, X_n$ is a submartingale with respect to $\mathscr{F}_1, \ldots, \mathscr{F}_n$ and τ_1, τ_2 are stopping times satisfying $1 \leq \tau_1 \leq \tau_2 \leq n$, then X_{τ_1}, X_{τ_2} is a submartingale with respect to $\mathscr{F}_{\tau_1}, \mathscr{F}_{\tau_2}$.*

This is the *optional sampling theorem*. The proof will show that X_{τ_1}, X_{τ_2} is a martingale if $X_1, \ldots, X_n$ is.

PROOF. Since the X_{τ_i} are dominated by $\sum_{k=1}^n |X_k|$, they are integrable. It is required to show that $E[X_{\tau_2} \| \mathscr{F}_{\tau_1}] \geq X_{\tau_1}$, or

$$(35.21) \qquad \int_A (X_{\tau_2} - X_{\tau_1}) \, dP \geq 0, \qquad A \in \mathscr{F}_{\tau_1}.$$

But $A \in \mathscr{F}_{\tau_1}$ implies that $A \cap [\tau_1 < k \leq \tau_2] = (A \cap [\tau_1 \leq k - 1]) \cap [\tau_2 \leq k - 1]^c$ lies in $\mathscr{F}_{k-1}$. If $\Delta_k = X_k - X_{k-1}$, then

$$\int_A (X_{\tau_2} - X_{\tau_1}) \, dP = \int_A \sum_{k=1}^n I_{[\tau_1 < k \leq \tau_2]} \Delta_k \, dP$$

$$= \sum_{k=1}^n \int_{A \cap [\tau_1 < k \leq \tau_2]} \Delta_k \, dP \geq 0$$

by the submartingale property. ∎

Inequalities

There are two inequalities that are fundamental to the theory of martingales.

Theorem 35.3. *If* $X_1, \ldots, X_n$ *is a submartingale, then for* $\alpha > 0$,

$$(35.22) \qquad P\left[\max_{i \leq n} X_i \geq \alpha \right] \leq \frac{1}{\alpha} E[|X_n|].$$

This extends Kolmogorov's inequality: If $S_1, S_2, \ldots$ are partial sums of independent random variables with mean 0, they form a martingale; if the variances are finite, then $S_1^2, S_2^2, \ldots$ is a submartingale by Theorem 35.1(i), and (35.22) for this submartingale is exactly Kolmogorov's inequality (22.9).

PROOF. Let $\tau_2 = n$; let τ_1 be the smallest k such that $X_k \geq \alpha$, if there is one, and n otherwise. If $M_k = \max_{i \leq k} X_i$, then $[M_n \geq \alpha] \cap [\tau_1 \leq k] = [M_k \geq \alpha] \in \mathscr{F}_k$, and hence $[M_n \geq \alpha]$ is in $\mathscr{F}_{\tau_1}$. By Theorem 35.2,

$$(35.23) \qquad \alpha P[M_n \geq \alpha] \leq \int_{[M_n \geq \alpha]} X_{\tau_1} \, dP \leq \int_{[M_n \geq \alpha]} X_n \, dP$$

$$\leq \int_{[M_n \geq \alpha]} X_n^+ \, dP \leq E[X_n^+] \leq E[|X_n|]. \quad ∎$$

This can also be proved by imitating the argument for Kolmogorov's inequality in Section 23. For improvements to (35.22), use the other integrals

in (35.23). If $X_1, \ldots, X_n$ is a martingale, $|X_1|, \ldots, |X_n|$ is a submartingale, and so (35.22) gives $P[\max_{i \le n} |X_i| \ge \alpha] \le \alpha^{-1} E[|X_n|]$.

The second fundamental inequality requires the notion of an *upcrossing*. Let $[\alpha, \beta]$ be an interval $(\alpha < \beta)$ and let $X_1, \ldots, X_n$ be random variables. Inductively define variables $\tau_1, \tau_2, \ldots$:

τ_1 is the smallest j such that $1 \le j \le n$ and $X_j \le \alpha$, and is n if there is no such j;

τ_k for *even* k is the smallest j such that $\tau_{k-1} < j \le n$ and $X_j \ge \beta$, and is n if there is no such j;

τ_k for *odd* k exceeding 1 is the smallest j such that $\tau_{k-1} < j \le n$ and $X_j \le \alpha$, and is n if there is no such j.

The number U of upcrossings of $[\alpha, \beta]$ by $X_1, \ldots, X_n$ is the largest i such that $X_{\tau_{2i-1}} \le \alpha < \beta \le X_{\tau_{2i}}$. In the diagram, $n = 20$ and there are three upcrossings.
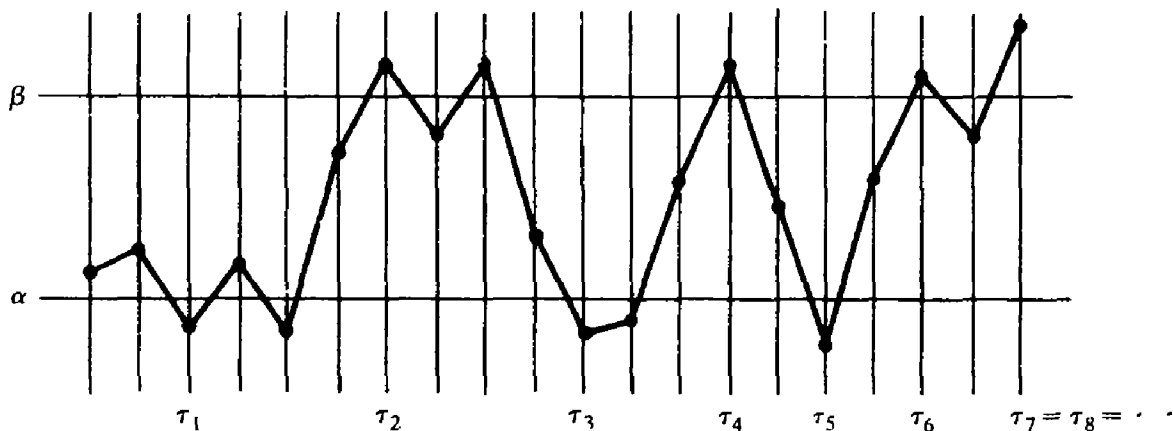
Theorem 35.4. *For a submartingale* $X_1, \ldots, X_n$, *the number* U *of upcrossings of* $[\alpha, \beta]$ *satisfies*

$$(35.24) \qquad E[U] \le \frac{E[|X_n|] + |\alpha|}{\beta - \alpha}.$$

PROOF. Let $Y_k = \max\{0, X_k - \alpha\}$ and $\theta = \beta - \alpha$. By Theorem 35.1(ii), $Y_1, \ldots, Y_n$ is a submartingale. The τ_k are unchanged if in the definitions $X_j \le \alpha$ is replaced by $Y_j = 0$ and $X_j \ge \beta$ by $Y_j \ge \theta$, and so U is also the number of upcrossings of $[0, \theta]$ by $Y_1, \ldots, Y_n$. If k is even and τ_{k-1} is a stopping time, then for $j < n$,

$$[\tau_k = j] = \bigcup_{i=1}^{j-1} [\tau_{k-1} = i, Y_{i+1} < \theta, \ldots, Y_{j-1} < \theta, Y_j \ge \theta]$$

lies in $\mathscr{F}_j$ and $[\tau_k = n] = [\tau_k \le n - 1]^c$ lies in $\mathscr{F}_n$, and so τ_k is also a stopping time. With a similar argument for odd k, this shows that the τ_k are all stopping times. Since the τ_k are strictly increasing until they reach n, $\tau_n = n$. Therefore,

$$Y_n = Y_{\tau_n} \ge Y_{\tau_n} - Y_{\tau_1} = \sum_{k=2}^{n} \left(Y_{\tau_k} - Y_{\tau_{k-1}} \right) = \Sigma_e + \Sigma_o,$$

where Σ_e and Σ_o are the sums over the even k and the odd k in the range $2 \le k \le n$. By Theorem 35.2, Σ_o has nonnegative expected value, and therefore, $E[Y_n] \ge E[\Sigma_e]$.

If $Y_{\tau_{2i-1}} = 0 < \theta \le Y_{\tau_{2i}}$ (which is the same thing as $X_{\tau_{2i-1}} \le \alpha < \beta \le X_{\tau_{2i}}$), then the difference $Y_{\tau_{2i}} - Y_{\tau_{2i-1}}$ appears in the sum Σ_e and is at least θ. Since there are U of these differences, $\Sigma_e \ge \theta U$, and therefore $E[Y_n] \ge \theta E[U]$. In terms of the original variables, this is

$$(\beta - \alpha) E[U] \le \int_{[X_n > \alpha]} (X_n - \alpha) \, dP \le E[|X_n|] + |\alpha|. \qquad \blacksquare$$

In a sense, an upcrossing of $[\alpha, \beta]$ is easy: since the X_k form a submartingale, they tend to increase. But before another upcrossing can occur, the sequence must make its way back down below α, which it resists. Think of the extreme case where the X_k are strictly increasing constants. This is reflected in the proof. Each of Σ_e and Σ_o has nonnegative expected value, but for Σ_e the proof uses the stronger inequality $E[\Sigma_e] \ge E[\theta U]$.

Convergence Theorems

The martingale convergence theorem, due to Doob, has a number of forms. The simplest one is this:

Theorem 35.5. *Let $X_1, X_2, \ldots$ be a submartingale. If $K = \sup_n E[|X_n|] < \infty$, then $X_n \to X$ with probability 1, where X is a random variable satisfying $E[|X|] \le K$.*

PROOF. Fix α and β for the moment, and let U_n be the number of upcrossings of $[\alpha, \beta]$ by $X_1, \ldots, X_n$. By the upcrossing theorem, $E[U_n] \le (E[|X_n|] + |\alpha|)/(\beta - \alpha) \le (K + |\alpha|)/(\beta - \alpha)$. Since U_n is nondecreasing and $E[U_n]$ is bounded, it follows by the monotone convergence theorem that $\sup_n U_n$ is integrable and hence finite-valued almost everywhere.

Let X^* and X_* be the limits superior and inferior of the sequence $X_1, X_2, \ldots$; they may be infinite. If $X_* < \alpha < \beta < X^*$, then U_n must go to infinity. Since U_n is bounded with probability 1, $P[X_* < \alpha < \beta < X^*] = 0$.

Now

$$(35.25) \qquad [X_* < X^*] = \bigcup [X_* < \alpha < \beta < X^*],$$

where the union extends over all pairs of rationals α and β. The set on the left therefore has probability 0.

Thus X^* and X_* are equal with probability 1, and X_n converges to their common value X, which may be $\pm\infty$. By Fatou's lemma, $E[|X|] \le \liminf_n E[|X_n|] \le K$. Since it is integrable, X is finite with probability 1. ∎

If the X_n form a martingale, then by (35.16) applied to the submartingale $|X_1|, |X_2|, \ldots$ the $E[|X_n|]$ are nondecreasing, so that $K = \lim_n E[|X_n|]$. The hypothesis in the theorem that K be finite is essential: If $X_n = \Delta_1 + \cdots + \Delta_n$, where the Δ_n are independent and assume values ± 1 with probability $\frac{1}{2}$, then X_n does not converge; $E[|X_n|]$ goes to infinity in this case.

If the X_n form a *nonnegative* martingale, then $E[|X_n|] = E[X_n] = E[X_1]$ by (35.5), and K is necessarily finite.

Example 35.9. The X_n in Example 35.6 are nonnegative, and so $X_n = Z_n / m^n \to X$, where X is nonnegative and integrable. If $m < 1$, then, since Z_n is an integer, $Z_n = 0$ for large n, and the population dies out. In this case, $X = 0$ with probability 1. Since $E[X_n] = E[X_0] = 1$, this shows that $E[X_n] \to E[X]$ may fail in Theorem 35.5. ∎

Theorem 35.5 has an important application to the martingale of Example 35.5, and this requires a lemma.

Lemma. *If Z is integrable and $\mathscr{F}_n$ are arbitrary σ-fields, then the random variables $E[Z\|\mathscr{F}_n]$ are uniformly integrable.*

For the definition of uniform integrability, see (16.21). The $\mathscr{F}_n$ must, of course, lie in the σ-field $\mathscr{F}$, but they need not, for example, be nondecreasing.

PROOF OF THE LEMMA. Since $|E[Z\|\mathscr{F}_n]| \le E[|Z| \,\|\mathscr{F}_n]$, Z may be assumed nonnegative. Let $A_{\alpha n} = [E[Z\|\mathscr{F}_n] \ge \alpha]$. Since $A_{\alpha n} \in \mathscr{F}_n$

$$\int_{A_{\alpha n}} E[Z\|\mathscr{F}_n]\, dP = \int_{A_{\alpha n}} Z\, dP.$$

It is therefore enough to find, for given ϵ, an α such that this last integral is less than ϵ for all n. Now $\int_A Z\, dP$ is, as a function of A, a finite measure dominated by P; by the ϵ-δ version of absolute continuity (see (32.4)) there is a δ such that $P(A) < \delta$ implies that $\int_A Z\, dP < \epsilon$. But $P[E[Z\|\mathscr{F}_n] \ge \alpha] \le \alpha^{-1} E[E[Z\|\mathscr{F}_n]] = \alpha^{-1} E[Z] < \delta$ for large enough α. ∎

Suppose that $\mathscr{F}_n$ are σ-fields satisfying $\mathscr{F}_1 \subset \mathscr{F}_2 \subset \cdots$. If the union $\bigcup_{n=1}^{\infty} \mathscr{F}_n$ generates the σ-field $\mathscr{F}_\infty$, this is expressed by $\mathscr{F}_n \uparrow \mathscr{F}_\infty$. The requirement is not that $\mathscr{F}_\infty$ coincide with the union, but that it be generated by it.

Theorem 35.6. *If $\mathscr{F}_n \uparrow \mathscr{F}_\infty$ and Z is integrable, then*

$$(35.26) \qquad\qquad E[Z\|\mathscr{F}_n] \to E[Z\|\mathscr{F}_\infty].$$

with probability 1.

PROOF.   According to Example 35.5, the random variables $X_n = E[Z\|\mathscr{F}_n]$ form a martingale relative to the $\mathscr{F}_n$. By the lemma, the X_n are uniformly integrable. Since $E[|X_n|] \le E[|Z|]$, by Theorem 35.5 the X_n converge to an integrable X. The problem is to identify X with $E[Z\|\mathscr{F}_\infty]$.

Because of the uniform integrability, it is possible (Theorem 16.14) to integrate to the limit: $\int_A X \, dP = \lim_n \int_A X_n \, dP$. If $A \in \mathscr{F}_k$ and $n \ge k$, then $\int_A X_n \, dP = \int_A E[Z\|\mathscr{F}_n] \, dP = \int_A Z \, dP$. Therefore, $\int_A X \, dP = \int_A Z \, dP$ for all A in the π-system $\bigcup_{k=1}^{\infty} \mathscr{F}_k$; since X is measurable $\mathscr{F}_\infty$, it follows by Theorem 34.1 that X is a version of $E[Z\|\mathscr{F}_\infty]$. ∎

Applications: Derivatives

Theorem 35.7.   *Suppose that $(\Omega, \mathscr{F}, P)$ is a probability space, ν is a finite measure on $\mathscr{F}$, and $\mathscr{F}_n \uparrow \mathscr{F}_\infty \subset \mathscr{F}$. Suppose that P dominates ν on each $\mathscr{F}_n$, and let X_n be the corresponding Radon–Nikodym derivatives. Then $X_n \to X$ with probability 1, where X is integrable.*

    (i)   *If P dominates ν on $\mathscr{F}_\infty$, then X is the corresponding Radon–Nikodym derivative.*

    (ii)   *If P and ν are mutually singular on $\mathscr{F}_\infty$, then $X = 0$ with probability 1.*

PROOF.   The situation is that of Example 35.2. The density X_n is measurable $\mathscr{F}_n$ and satisfies (35.9). Since X_n is nonnegative, $E[|X_n|] = E[X_n] = \nu(\Omega)$, and it follows by Theorem 35.5 that X_n converges to an integrable X. The limit X is measurable $\mathscr{F}_\infty$.

Suppose that P dominates ν on $\mathscr{F}_\infty$ and let Z be the Radon–Nikodym derivative: Z is measurable $\mathscr{F}_\infty$, and $\int_A Z \, dP = \nu(A)$ for $A \in \mathscr{F}_\infty$. It follows that $\int_A Z \, dP = \int_A X_n \, dP$ for A in $\mathscr{F}_n$, and so $X_n = E[Z\|\mathscr{F}_n]$. Now Theorem 35.6 implies that $X_n \to E[Z\|\mathscr{F}_\infty] = Z$.

Suppose, on the other hand, that P and ν are mutually singular on $\mathscr{F}_\infty$, so that there exists a set S in $\mathscr{F}_\infty$ such that $\nu(S) = 0$ and $P(S) = 1$. By Fatou's lemma $\int_A X \, dP \le \liminf_n \int_A X_n \, dP$. If $A \in \mathscr{F}_k$, then $\int_A X_n \, dP = \nu(A)$ for $n \ge k$, and so $\int_A X \, dP \le \nu(A)$ for A in the field $\bigcup_{k=1}^{\infty} \mathscr{F}_k$. It follows by the monotone class theorem that this holds for all A in $\mathscr{F}_\infty$. Therefore, $\int X \, dP = \int_S X \, dP \le \nu(S) = 0$, and X vanishes with probability 1. ∎

Example 35.10. As in Example 35.3, let ν be a finite measure on the unit interval with Lebesgue measure $(\Omega, \mathscr{F}, P)$. For $\mathscr{F}_n$ the σ-field generated by the dyadic intervals of rank n, (35.10) gives X_n. In this case $\mathscr{F}_n \uparrow \mathscr{F}_\infty = \mathscr{F}$. For each ω and n choose the dyadic rationals $a_n(\omega) = k2^{-n}$ and $b_n(\omega) = (k+1)2^{-n}$ for which $a_n(\omega) < \omega \le b_n(\omega)$. By Theorem 35.7, if F is the distribution function for ν, then

$$(35.27) \qquad \frac{F(b_n(\omega)) - F(a_n(\omega))}{b_n(\omega) - a_n(\omega)} \to X(\omega)$$

except on a set of Lebesgue measure 0.

According to Theorem 31.2, F has a derivative F' except on a set of Lebesgue measure 0, and since the intervals $(a_n(\omega), b_n(\omega)]$ contract to ω, the difference ratio (35.27) converges almost everywhere to $F'(\omega)$ (see (31.8)). This identifies X. Since (35.27) involves intervals $(a_n(\omega), b_n(\omega)]$ of a special kind, it does not quite imply Theorem 31.2.

By Theorem 35.7, $X = F'$ is the density for ν in the absolutely continuous case, and $X = F' = 0$ (except on a set of Lebesgue measure 0) in the singular case, facts proved in a different way in Section 31. The singular case gives another example where $E[X_n] \to E[X]$ fails in Theorem 35.5. ∎

Likelihood Ratios

Return to Example 35.4: $\nu = Q$ is a probability measure, $\mathscr{F}_n = \sigma(Y_1, \ldots, Y_n)$ for random variables Y_n, and the Radon–Nikodym derivative or likelihood ratio X_n has the form (35.11) for densities p_n and q_n on R^n. By Theorem 35.7 the X_n converge to some X which is integrable and measurable $\mathscr{F}_\infty = \sigma(Y_1, Y_2, \ldots)$.

If the Y_n are independent under P and under Q, and if the densities are different, then P and Q are mutually singular on $\sigma(Y_1, Y_2, \ldots)$, as shown in Example 35.4. In this case $X = 0$ and the likelihood ratio converges to 0 on a set of P-measure 1. The statistical relevance of this is that the smaller X_n is, the more strongly one prefers P over Q as an explanation of the observation $(Y_1, \ldots, Y_n)$, and X_n goes to 0 with probability 1 if P is in fact the measure governing the Y_n.

It might be thought that a disingenuous experimenter could bias his results by stopping at an X_n he likes—a large value if his prejudices favor Q, a small value if they favor P. This is not so, as the following analysis shows. For this argument P must dominate Q on each $\mathscr{F}_n = \sigma(Y_1, \ldots, Y_n)$, but the likelihood ratio X_n need not have any special form.

Let τ be a positive-integer-valued random variable representing the time the experimenter stops. Assume that τ is finite, and to exclude prevision, assume that it is a stopping time. The σ-field $\mathscr{F}_\tau$ defined by (35.20) represents the information available at time τ, and the problem is to show that X_τ

is the likelihood ratio (Radon–Nikodym derivative) for Q with respect to P on $\mathcal{F}_\tau$. First, X_τ is clearly measurable $\mathcal{F}_\tau$. Second, if $A \in \mathcal{F}_\tau$, then $A \cap [\tau = n] \in \mathcal{F}_n$, and therefore

$$\int_A X_\tau \, dP = \sum_{n=1}^\infty \int_{A \cap [\tau = n]} X_n \, dP = \sum_{n=1}^\infty Q(A \cap [\tau = n]) = Q(A),$$

as required.

Reversed Martingales

A left-infinite sequence $\ldots, X_{-2}, X_{-1}$ is a martingale relative to σ-fields $\ldots, \mathcal{F}_{-2}, \mathcal{F}_{-1}$ if conditions (ii) and (iii) in the definition of martingale are satisfied for $n \leq -1$ and conditions (i) and (iv) are satisfied for $n < -1$. Such a sequence is a *reversed* or *backward* martingale.

Theorem 35.8. *For a reversed martingale, $\lim_{n \to \infty} X_{-n} = X$ exists and is integrable, and $E[X] = E[X_{-n}]$ for all n.*

PROOF. The proof is almost the same as that for Theorem 35.5. Let X^* and X_* be the limits superior and inferior of the sequence $X_{-1}, X_{-2}, \ldots$. Again (35.25) holds. Let U_n be the number of upcrossings of $[\alpha, \beta]$ by $X_{-n}, \ldots, X_{-1}$. By the upcrossing theorem, $E[U_n] \leq (E[|X_{-1}|] + |\alpha|)/(\beta - \alpha)$. Again $E[U_n]$ is bounded, and so $\sup_n U_n$ is finite with probability 1 and the sets in (35.25) have probability 0.

Therefore, $\lim_{n \to \infty} X_{-n} = X$ exists with probability 1. By the property (35.4) for martingales, $X_{-n} = E[X_{-1} \| \mathcal{F}_{-n}]$ for $n = 1, 2, \ldots$. The lemma above (p. 469) implies that the X_{-n} are uniformly integrable. Therefore, X is integrable and $E[X]$ is the limit of the $E[X_{-n}]$; these all have the same value by (35.5). ∎

If $\mathcal{F}_n$ are σ-fields satisfying $\mathcal{F}_1 \supset \mathcal{F}_2 \ldots$, then the intersection $\bigcap_{n=1}^\infty \mathcal{F}_n = \mathcal{F}_0$ is also a σ-field, and the relation is expressed by $\mathcal{F}_n \downarrow \mathcal{F}_0$.

Theorem 35.9. *If $\mathcal{F}_n \downarrow \mathcal{F}_0$ and Z is integrable, then*

$$(35.28) \qquad\qquad E[Z \| \mathcal{F}_n] \to E[Z \| \mathcal{F}_0]$$

with probability 1.

PROOF. If $X_{-n} = E[Z \| \mathcal{F}_n]$, then $\ldots, X_{-2}, X_{-1}$ is a martingale relative to $\ldots, \mathcal{F}_2, \mathcal{F}_1$. By the preceding theorem, $E[Z \| \mathcal{F}_n]$ converges as $n \to \infty$ to an integrable X and by the lemma, the $E[Z \| \mathcal{F}_n]$ are uniformly integrable. As the limit of the $E[Z \| \mathcal{F}_n]$ for $n \geq k$, X is measurable $\mathcal{F}_k$; k being arbitrary, X is measurable $\mathcal{F}_0$.

By uniform integrability, $A \in \mathscr{F}_0$ implies that

$$\int_A X \, dP = \lim_n \int_A E[Z\|\mathscr{F}_n] \, dP = \lim_n \int_A E[E[Z\|\mathscr{F}_n]\|\mathscr{F}_0] \, dP$$

$$= \lim_n \int_A E[Z\|\mathscr{F}_0] \, dP = \int_A E[Z\|\mathscr{F}_0] \, dP.$$

Thus X is a version of $E[Z\|\mathscr{F}_0]$.   ∎

Theorems 35.6 and 35.9 are parallel. There is an essential difference between Theorems 35.5 and 35.8, however. In the latter, the martingale has a last random variable, namely X_{-1}, and so it is unnecessary in proving convergence to assume the $E[|X_n|]$ bounded. On the other hand, the proof in Theorem 35.8 that X is integrable would not work for a submartingale.

Applications: de Finetti's Theorem

Let $\theta, X_1, X_2, \ldots$ be random variables such that $0 \le \theta \le 1$ and, conditionally on θ, the X_n are independent and assume the values 1 and 0 with probabilities θ and $1 - \theta$: for $u_1, \ldots, u_n$ a sequence of 0's and 1's,

$$(35.29) \qquad P[X_1 = u_1, \ldots, X_n = u_n \| \theta] = \theta^s (1 - \theta)^{n-s},$$

where $s = u_1 + \cdots + u_n$.

To see that such sequences exist, let $\theta, Z_1, Z_2, \ldots$ be independent random variables, where θ has an arbitrarily prescribed distribution supported by $[0, 1]$ and the Z_n are uniformly distributed over $[0, 1]$. Put $X_n = I_{[Z_n \le \theta]}$. If, for instance, $f(x) = x(1 - x) = P[Z_1 \le x, Z_2 > x]$, then $P[X_1 = 1, X_2 = 0\|\theta]_\omega = f(\theta(\omega))$ by (33.13). The obvious extension establishes (35.29).

Integrate (35.29):

$$(35.30) \qquad P[X_1 = u_1, \ldots, X_n = u_n] = E[\theta^s (1 - \theta)^{n-s}].$$

Thus $\{X_n\}$ is a mixture of Bernoulli processes. It is clear from (35.30) that the X_k are *exchangeable* in the sense that for each n the distribution of $(X_1, \ldots, X_n)$ is invariant under permutations. According to the following theorem of de Finetti, *every* exchangeable sequence is a mixture of Bernoulli sequences.

Theorem 35.10. *If the random variables $X_1, X_2, \ldots$ are exchangeable and take values 0 and 1, then there is a random variable θ for which (35.29) and (35.30) hold.*

PROOF. Let $S_m = X_1 + \cdots + X_m$. If $t \leq m$, then

$$P[S_m = t] = \sum_{u_1 \quad u_m} P[X_1 = u_1, \ldots, X_m = u_m],$$

where the sum extends over the sequences for which $u_1 + \cdots + u_m = t$. By exchangeability, the terms on the right are all equal, and since there are $\binom{m}{t}$ of them,

$$P[X_1 = u_1, \ldots, X_m = u_m | S_m = t] = \binom{m}{t}^{-1}.$$

Suppose that $s \leq n \leq m$ and $u_1 + \cdots + u_n = s \leq t \leq m$; add out the $u_{n+1}, \ldots, u_m$ that sum to $t - s$:

$$P[X_1 = u_1, \ldots, X_n = u_n | S_m = t] = \binom{m-n}{t-s} \bigg/ \binom{m}{t}$$

$$= \frac{(t)_s (m-t)_{n-s}}{(m)_n} = f_{n,s,m}\left(\frac{t}{m}\right),$$

where

$$f_{n,s,m}(x) = \prod_{i=0}^{s-1}\left(x - \frac{i}{m}\right) \prod_{i=0}^{n-s-1}\left(1 - x - \frac{i}{m}\right) \bigg/ \prod_{i=0}^{n-1}\left(1 - \frac{i}{m}\right).$$

The preceding equations still hold if further constraints $S_{m+1} = t_1, \ldots, S_{m+j} = t_j$ are joined to $S_m = t$. Therefore, $P[X_1 = u_1, \ldots, X_n = u_n \| S_m, \ldots, S_{m+j}] = f_{n,s,m}(S_m/m)$.

Let $\mathcal{S}_m = \sigma(S_m, S_{m+1}, \ldots)$ and $\mathcal{S} = \bigcap_m \mathcal{S}_m$. Now fix n and $u_1, \ldots, u_n$, and suppose that $u_1 + \cdots + u_n = s$. Let $j \to \infty$ and apply Theorem 35.6: $P[X_1 = u_1, \ldots, X_n = u_n \| \mathcal{S}_m] = f_{n,s,m}(S_m/m)$. Let $m \to \infty$ and apply Theorem 35.9:

$$P[X_1 = u_1, \ldots, X_n = u_n \| \mathcal{S}]_\omega = \lim_m f_{n,s,m}\left(\frac{S_m(\omega)}{m}\right)$$

holds for ω outside a set of probability 0.

Fix such an ω and suppose that $\{S_m(\omega)/m\}$ has two distinct limit points. Since the distance from each $S_m(\omega)/m$ to the next is less than $2/m$, it follows that the set of limit points must fill a nondegenerate interval. But $\lim_k x_{m_k} = x$ implies $\lim_k f_{n,s,m_k}(x_{m_k}) = x^s(1-x)^{n-s}$, and so it follows further that $x^s(1-x)^{n-s}$ must be constant over this interval, which is impossible. Therefore, $S_m(\omega)/m$ must converge to some limit $\theta(\omega)$. This shows that $P[X_1 = u_1, \ldots, X_n = u_n \| \mathcal{S}] = \theta^s(1-\theta)^{n-s}$ with probability 1. Take a conditional expectation with respect to $\sigma(\theta)$, and (35.29) follows. ∎

Bayes Estimation

From the Bayes point of view in statistics, the θ in (35.29) is a parameter governed by some a priori distribution known to the statistician. For given $X_1, \ldots, X_n$, the Bayes estimate of θ is $E[\theta \| X_1, \ldots, X_n]$. The problem is to show that this estimate is consistent in the sense that

$$(35.31) \qquad\qquad E[\theta \| X_1, \ldots, X_n] \to \theta$$

with probability 1. By Theorem 35.6, $E[\theta \| X_1, \ldots, X_n] \to E[\theta \| \mathscr{F}_\infty]$, where $\mathscr{F}_\infty = \sigma(X_1, X_2, \ldots)$, and so what must be shown is that $E[\theta \| \mathscr{F}_\infty] = \theta$ with probability 1.

By an elementary argument that parallels the unconditional case, it follows from (35.29) for $S_n = X_1 + \cdots + X_n$ that $E[S_n \| \theta] = n\theta$ and $E[(S_n - n\theta)^2 \| \theta] = n\theta(1 - \theta)$. Hence $E[(n^{-1}S_n - \theta)^2] \le n^{-1}$, and by Chebyshev's inequality $n^{-1}S_n$ converges in probability to θ. Therefore (Theorem 20.5), $\lim_k n_k^{-1} S_{n_k} = \theta$ with probability 1 for some subsequence. Thus $\theta = \theta'$ with probability 1 for a θ' measurable $\mathscr{F}_\infty$, and $E[\theta \| \mathscr{F}_\infty] = E[\theta' \| \mathscr{F}_\infty] = \theta' = \theta$ with probability 1.

A Central Limit Theorem*

Suppose $X_1, X_2, \ldots$ is a martingale relative to $\mathscr{F}_1, \mathscr{F}_2, \ldots$, and put $Y_n = X_n - X_{n-1}$ ($Y_1 = X_1$), so that

$$(35.32) \qquad\qquad E[Y_n \| \mathscr{F}_{n-1}] = 0.$$

View Y_n as the gambler's gain on the nth trial in a fair game. For example, if $\Delta_1, \Delta_2, \ldots$ are independent and have mean 0, $\mathscr{F}_n = \sigma(\Delta_1, \ldots, \Delta_n)$, W_n is measurable $\mathscr{F}_{n-1}$, and $Y_n = W_n \Delta_n$, then (35.32) holds (see (35.17)). A specialization of this case shows that $X_n = \sum_{k=1}^n Y_k$ need not be approximately normally distributed for large n.

Example 35.11. Suppose that Δ_n takes the values ± 1 with probability $\frac{1}{2}$ each and $W_1 = 0$, and suppose that $W_n = 1$ for $n \ge 2$ if $\Delta_1 = 1$, while $W_n = 2$ for $n \ge 2$ if $\Delta_1 = -1$. If $S_n = \Delta_2 + \cdots + \Delta_n$, then X_n is S_n or $2S_n$ according as Δ_1 is $+1$ or -1. Since $S_n / \sqrt{n}$ has approximately the standard normal distribution, the approximate distribution of $X_n / \sqrt{n}$ is a mixture, with equal weights, of the centered normal distributions with standard deviations 1 and 2. ∎

To understand this phenomenon, consider conditional variances. Suppose for simplicity that the Y_n are bounded, and define

$$(35.33) \qquad\qquad \sigma_n^2 = E[Y_n^2 \| \mathscr{F}_{n-1}]$$

---

*This topic, which requires the limit theory of Chapter 5, may be omitted.

(take $\mathscr{F}_0 = \{\varnothing, \Omega\}$). Consider the stopping times

$$(35.34) \qquad \nu_t = \min\left[n: \sum_{k=1}^{n} \sigma_k^2 \ge t \right].$$

Under appropriate conditions, $X_{\nu_t}/\sqrt{t}$ will be approximately normally distributed for large t. Consider the preceding example. Roughly: If $\Delta_1 = +1$, then $\sum_{k=1}^{n} \sigma_k^2 = n - 1$, and so $\nu_t \approx t$ and $X_{\nu_t}/\sqrt{t} \approx S_t/\sqrt{t}$; if $\Delta_1 = -1$, then $\sum_{k=1}^{n} \sigma_k^2 = 4(n - 1)$, and so $\nu_t \approx t/4$ and $X_{\nu_t}/\sqrt{t} \approx 2S_{t/4}/\sqrt{t} = S_{t/4}/\sqrt{t/4}$. In either case, $X_{\nu_t}/\sqrt{t}$ approximately follows the standard normal law.

If the nth play takes σ_n^2 units of time, then ν_t is essentially the number of plays that take place during the first t units of time. This change of the time scale stabilizes the rate at which money changes hands.

Theorem 35.11. *Suppose the $Y_n = X_n - X_{n-1}$ are uniformly bounded and satisfy (35.32), and assume that $\sum_n \sigma_n^2 = \infty$ with probability 1. Then $X_{\nu_t}/\sqrt{t} \Rightarrow N$.*

This will be deduced from a more general result, one that contains the Lindeberg theorem. Suppose that, for each n, $X_{n1}, X_{n2}, \ldots$ is a martingale with respect to $\mathscr{F}_{n1}, \mathscr{F}_{n2}, \ldots$. Define $Y_{nk} = X_{nk} - X_{n,k-1}$, suppose the Y_{nk} have second moments, and put $\sigma_{nk}^2 = E[Y_{nk}^2 \| \mathscr{F}_{n,k-1}](\mathscr{F}_{n0} = \{\varnothing, \Omega\})$. The probability space may vary with n. If the martingale is originally defined only for $1 \le k \le r_n$, take $Y_{nk} = 0$ and $\mathscr{F}_{nk} = \mathscr{F}_{nr_n}$ for $k > r_n$. Assume that $\sum_{k=1}^{\infty} Y_{nk}$ and $\sum_{k=1}^{\infty} \sigma_{nk}^2$ converge with probability 1.

Theorem 35.12. *Suppose that*

$$(35.35) \qquad \sum_{k=1}^{\infty} \sigma_{nk}^2 \to_P \sigma^2,$$

where σ is a positive constant, and that

$$(36.36) \qquad \sum_{k=1}^{\infty} E\left[Y_{nk}^2 I_{[|Y_{nk}| \ge \epsilon]} \right] \to 0$$

for each ϵ. Then $\sum_{k=1}^{\infty} Y_{nk} \Rightarrow \sigma N$.

PROOF OF THEOREM 35.11.   The proof will be given for t going to infinity through the integers.[†] Let $Y_{nk} = I_{[\nu_n \ge k]} Y_k/\sqrt{n}$ and $\mathscr{F}_{nk} = \mathscr{F}_k$. From $[\nu_n \ge k]$ $= [\sum_{j=1}^{k-1} \sigma_j^2 < n] \in \mathscr{F}_{k-2}$ follow $E[Y_{nk} \| \mathscr{F}_{n,k-1}] = 0$ and $\sigma_{nk}^2 = E[Y_{nk}^2 \| \mathscr{F}_{n,k-1}]$ $= I_{[\nu_n \ge k]} \sigma_k^2/n$. If K bounds the $|Y_k|$, then $1 \le \sum_{k=1}^{\infty} \sigma_{nk}^2 = n^{-1} \sum_{k=1}^{\nu_n} \sigma_k^2 \le 1 + K^2/n$, so that (35.35) holds for $\sigma = 1$. For n large enough that $K/\sqrt{n} < \epsilon$, the sum in (35.36) vanishes. Theorem 35.12 therefore applies, and $\sum_{k=1}^{\nu_n} Y_k/\sqrt{n} = \sum_{k=1}^{\infty} Y_{nk} \Rightarrow N$.   ∎

[†]For the general case, first check that the proof of Theorem 35.12 goes through without change if n is replaced by a parameter going continuously to infinity.

PROOF OF THEOREM 35.12. Assume at first that there is a constant c such that

$$(35.37) \qquad \sum_{k=1}^{\infty} \sigma_{nk}^2 \leq c,$$

which in fact suffices for the application to Theorem 35.11.

Write $S_k = \sum_{j=1}^{k} Y_{nj}$ ($S_0 = 0$), $S_\infty = \sum_{j=1}^{\infty} Y_{nj}$, $\Sigma_k = \sum_{j=1}^{k} \sigma_{nj}^2$ ($\Sigma_0 = 0$), and $\Sigma_\infty = \sum_{j=1}^{\infty} \sigma_{nj}^2$; the dependence on n is suppressed in the notation. To prove $E[e^{itS_x}] \to e^{-\frac{1}{2}t^2\sigma^2}$, observe first that

$$\left| E\left[e^{itS_x} - e^{-\frac{1}{2}t^2\sigma^2} \right] \right|$$

$$= \left| E\left[e^{itS_x}(1 - e^{\frac{1}{2}t^2\Sigma_x} e^{-\frac{1}{2}t^2\sigma^2}) + e^{-\frac{1}{2}t^2\sigma^2}(e^{\frac{1}{2}t^2\Sigma_x} e^{itS_x} - 1) \right] \right|$$

$$\leq E\left[|1 - e^{\frac{1}{2}t^2\Sigma_x} e^{-\frac{1}{2}t^2\sigma^2}| \right] + \left| E\left[e^{\frac{1}{2}t^2\Sigma_x} e^{itS_x} - 1 \right] \right| = A + B.$$

The term A on the right goes to 0 as $n \to \infty$, because by (35.35) and (35.37) the integrand is bounded and goes to 0 in probability.

The integrand in B is

$$\sum_{k=1}^{\infty} e^{itS_{k-1}}(e^{itY_{r,k}} - e^{-\frac{1}{2}t^2\sigma_{nk}^2}) e^{\frac{1}{2}t^2\Sigma_k},$$

because the mth partial sum here telescopes to $e^{itS_m} e^{\frac{1}{2}t^2\Sigma_m} - 1$. Since, by (35.37), this partial sum is bounded uniformly in m, and since S_{k-1} and Σ_k are measurable $\mathscr{F}_{r,k-1}$, it follows (Theorem 16.7) that

$$B = \left| \sum_{k=1}^{\infty} E\left[e^{itS_{k-1}} e^{\frac{1}{2}t^2\Sigma_k}(e^{itY_{nk}} - e^{-\frac{1}{2}t^2\sigma_{nk}^2}) \right] \right|$$

$$\leq \sum_{k=1}^{\infty} \left| E\left[e^{itS_{k-1}} e^{\frac{1}{2}t^2\Sigma_k} E\left[e^{itY_{nk}} - e^{-\frac{1}{2}t^2\sigma_{nk}^2} \| \mathscr{F}_{n,k-1} \right] \right] \right|$$

$$\leq e^{\frac{1}{2}t^2c} \sum_{k=1}^{\infty} E\left[\left| E\left[e^{itY_{nk}} - e^{-\frac{1}{2}t^2\sigma_{nk}^2} \| \mathscr{F}_{n,k-1} \right] \right| \right].$$

To complete the proof (under the temporary assumption (35.37)), it is enough to show that this last sum goes to 0.

By (26.4$_2$),

$$(35.38) \qquad e^{itY_{nk}} = 1 + itY_{nk} - \tfrac{1}{2}t^2 Y_{nk}^2 + \theta,$$

where (write $I_{nk} = I_{[|Y_{nk}| \geq \epsilon]}$ and let K_t bound t^2 and $|t|^3$)

$$|\theta| \leq \min\{|tY_{nk}|^3, |tY_{nk}|^2\} \leq K_t(Y_{nk}^2 I_{nk} + \epsilon Y_{nk}^2).$$

And

(38.39) $$e^{-\frac{1}{2}t^2\sigma_{nk}^2} = 1 - \tfrac{1}{2}t^2\sigma_{nk}^2 + \theta',$$

where (use (27.15) and increase K_t)

$$|\theta'| \le \left(\tfrac{1}{2}t^2\sigma_{nk}^2\right)^2 e^{\frac{1}{2}t^2\sigma_{nk}^2} \le t^4\sigma_{nk}^4 e^{\frac{1}{2}t^2 c} \le K_t\sigma_{nk}^4.$$

Because of the condition $E[Y_{nk}\|\mathscr{F}_{n,k-1}] = 0$ and the definition of σ_{nk}^2, the right sides of (35.38) and (35.39), minus θ and θ', respectively, have the same conditional expected value given $\mathscr{F}_{n,k-1}$. By (35.37), therefore,

$$\sum_{k=1}^{\infty} E\left[\left|E\left[e^{itY_{nk}} - e^{-\frac{1}{2}t^2\sigma_{nk}^2}\|\mathscr{F}_{n,k-1}\right]\right|\right]$$

$$\le K_t \sum_{k=1}^{\infty} \left(E\left[Y_{nk}^2 I_{nk}\right] + \epsilon E\left[\sigma_{nk}^2\right] + E\left[\sigma_{nk}^4\right]\right)$$

$$\le K_t \left(\sum_{k=1}^{\infty} E\left[Y_{nk}^2 I_{nk}\right] + \epsilon c + cE\left[\sup_{k\ge 1}\sigma_{nk}^2\right]\right).$$

Since $\sigma_{nk}^2 \le E[\epsilon^2 + Y_{nk}^2 I_{nk}\|\mathscr{F}_{n,k-1}] \le \epsilon^2 + \sum_{j=1}^{\infty}E[Y_{nj}^2 I_{nj}\|\mathscr{F}_{n,j-1}]$, it follows by (35.36) that the last expression above is, in the limit, at most $K_t(\epsilon c + c\epsilon^2)$. Since ϵ is arbitrary, this completes the proof of the theorem under the assumption (35.37).

To remove this assumption, take $c > \sigma^2$, define $A_{nk} = [\sum_{j=1}^{k}\sigma_{nj}^2 \le c]$ and $A_{n\infty} = [\sum_{j=1}^{\infty}\sigma_{nj}^2 \le c]$, and take $Z_{nk} = Y_{nk}I_{A_{nk}}$. From $A_{nk} \in \mathscr{F}_{n,k-1}$ follow $E[Z_{nk}\|\mathscr{F}_{n,k-1}] = 0$ and $\tau_{nk}^2 = E[Z_{nk}^2\|\mathscr{F}_{n,k-1}] = I_{A_{nk}}\sigma_{nk}^2$. Since $\sum_{j=1}^{\infty}\tau_{nj}^2$ is $\sum_{j=1}^{k}\sigma_{nj}^2$ on $A_{nk} - A_{n,k+1}$ and $\sum_{j=1}^{\infty}\sigma_{nj}^2$ on $A_{n\infty}$, the Z-array satisfies (35.37). Now $P(A_{n\infty}) \to 1$ by (35.35), and on $A_{n\infty}$, $\tau_{nk}^2 = \sigma_{nk}^2$ for all k, so that the Z-array satisfies (35.35). And it satisfies (35.36) because $|Z_{nk}| \le |Y_{nk}|$. Therefore, by the case already treated, $\sum_{k=1}^{\infty}Z_{nk} \Rightarrow \sigma N$. But since $\sum_{k=1}^{\infty}Y_{nk}$ coincides with this last sum on $A_{n\infty}$, it, too, is asymptotically normal. ∎

PROBLEMS

35.1. Suppose that $\Delta_1, \Delta_2, \ldots$ are independent random variables with mean 0. Let $X_1 = \Delta_1$ and $X_{n+1} = X_n + \Delta_{n+1}f_n(X_1,\ldots,X_n)$, and suppose that the X_n are integrable. Show that $\{X_n\}$ is a martingale. The martingales of gambling have this form.

35.2. Let $Y_1, Y_2, \ldots$ be independent random variables with mean 0 and variance σ^2. Let $X_n = (\sum_{k=1}^{n}Y_k)^2 - n\sigma^2$ and show that $\{X_n\}$ is a martingale.

35.3. Suppose that $\{Y_n\}$ is a finite-state Markov chain with transition matrix $[p_{ij}]$. Suppose that $\sum_j p_{ij} x(j) = \lambda x(i)$ for all i (the $x(i)$ are the components of a right eigenvector of the transition matrix). Put $X_n = \lambda^{-n} x(Y_n)$ and show that $\{X_n\}$ is a martingale.

35.4. Suppose that $Y_1, Y_2, \ldots$ are independent, positive random variables and that $E[Y_n] = 1$. Put $X_n = Y_1 \cdots Y_n$.

(a) Show that $\{X_n\}$ is a martingale and converges with probability 1 to an integrable X.

(b) Suppose specifically that Y_n assumes the values $\frac{1}{2}$ and $\frac{3}{2}$ with probability $\frac{1}{2}$ each. Show that $X = 0$ with probability 1. This gives an example where $E[\prod_{n=1}^{\infty} Y_n] \neq \prod_{n=1}^{\infty} E[Y_n]$ for independent, integrable, positive random variables. Show, however, that $E[\prod_{n=1}^{\infty} Y_n] \leq \prod_{n=1}^{\infty} E[Y_n]$ always holds.

35.5. Suppose that $X_1, X_2, \ldots$ is a martingale satisfying $E[X_1] = 0$ and $E[X_n^2] < \infty$. Show that $E[(X_{n+r} - X_n)^2] = \sum_{k=1}^{r} E[(X_{n+k} - X_{n+k-1})^2]$ (the variance of the sum is the sum of the variances). Assume that $\sum_n E[(X_n - X_{n-1})^2] < \infty$ and prove that X_n converges with probability 1. Do this first by Theorem 35.5 and then (see Theorem 22.6) by Theorem 35.3.

35.6. Show that a submartingale X_n can be represented as $X_n = Y_n + Z_n$, where Y_n is a martingale and $0 \leq Z_1 \leq Z_2 \leq \cdots$. *Hint*: Take $X_0 = 0$ and $\Delta_n = X_n - X_{n-1}$, and define $Z_n = \sum_{k=1}^{n} E[\Delta_k \| \mathscr{F}_{k-1}]$ ($\mathscr{F}_0 = \{0, \Omega\}$).

35.7. If $X_1, X_2, \ldots$ is a martingale and bounded either above or below, then $\sup_n E[|X_n|] < \infty$.

35.8. $\uparrow$ Let $X_n = \Delta_1 + \cdots + \Delta_n$, where the Δ_n are independent and assume the values ± 1 with probability $\frac{1}{2}$ each. Let τ be the smallest n such that $X_n = 1$ and define X_n^* by (35.19). Show that the hypotheses of Theorem 35.5 are satisfied by $\{X_n^*\}$ but that it is impossible to integrate to the limit. *Hint*: Use (7.8) and Problem 35.7.

35.9. Let $X_1, X_2, \ldots$ be a martingale, and assume that $|X_1(\omega)|$ and $|X_n(\omega) - X_{n-1}(\omega)|$ are bounded by a constant independent of ω and n. Let τ be a stopping time with finite mean. Show that X_τ is integrable and that $E[X_\tau] = E[X_1]$.

35.10. 35.8 35.9 $\uparrow$ Use the preceding result to show that the τ in Problem 35.8 has infinite mean. Thus the waiting time until a symmetric random walk moves one step up from the starting point has infinite expected value.

35.11. Let $X_1, X_2, \ldots$ be a Markov chain with countable state space S and transition probabilities p_{ij}. A function φ on S is excessive or superharmonic if $\varphi(i) \geq \sum_j p_{ij} \varphi(j)$. Show by martingale theory that $\varphi(X_n)$ converges with probability 1 if φ is bounded and excessive. Deduce from this that if the chain is irreducible and persistent, then φ must be constant. Compare Problem 8.34.

35.12. $\uparrow$ A function φ on the integer lattice in R^k is superharmonic if for each lattice point x, $\varphi(x) \geq (2k)^{-1} \sum \varphi(y)$, the sum extending over the $2k$ nearest neighbors y. Show for $k = 1$ and $k = 2$ that a bounded superharmonic function is constant. Show for $k \geq 3$ that there exist nonconstant bounded harmonic functions.

35.13. **32.7** **32.9↑** Let $(\Omega, \mathcal{F}, P)$ be a probability space, let ν be a finite measure on $\mathcal{F}$, and suppose that $\mathcal{F}_n \uparrow \mathcal{F}_\infty \subset \mathcal{F}$. For $n \le \infty$, let X_n be the Radon–Nikodym derivative with respect to P of the absolutely continuous part of ν when P and ν are both restricted to $\mathcal{F}_n$. The problem is to extend Theorem 35.7 by showing that $X_n \to X_\infty$ with probability 1.

(a) For $n \le \infty$, let

$$\nu(A) = \int_A X_n \, dP + \sigma_n(A), \qquad A \in \mathcal{F}_n,$$

be the decomposition of ν into absolutely continuous and singular parts with respect to P on $\mathcal{F}_n$. Show that $X_1, X_2, \ldots$ is a supermartingale and converges with probability 1.

(b) Let

$$\sigma_\infty(A) = \int_A Z_n \, dP + \sigma_n'(A), \qquad A \in \mathcal{F}_n,$$

be the decomposition of σ_∞ into absolutely continuous and singular parts with respect to P on $\mathcal{F}_n$. Let $Y_n = E[X_\infty \| \mathcal{F}_n]$, and prove

$$\int_A (Y_n + Z_n) \, dP + \sigma_n'(A) = \int_A X_n \, dP + \sigma_n(A), \qquad A \in \mathcal{F}_n.$$

Conclude that $Y_n + Z_n = X_n$ with probability 1. Since Y_n converges to X_∞, Z_n converges with probability 1 to some Z. Show that $\int_A Z \, dP \le \sigma_\infty(A)$ for $A \in \mathcal{F}_\infty$, and conclude that $Z = 0$ with probability 1.

35.14. (a) Show that $\{X_n\}$ is a martingale with respect to $\{\mathcal{F}_n\}$ if and only if, for all n and all stopping times τ such that $\tau \le n$, $E[X_n \| \mathcal{F}_\tau] = X_\tau$.

(b) Show that, if $\{X_n\}$ is a martingale and τ is a bounded stopping time, then $E[X_\tau] = E[X_1]$.

35.15. **31.9↑** Suppose that $\mathcal{F}_n \uparrow \mathcal{F}_\infty$ and $A \in \mathcal{F}_\infty$ and prove that $P[A \| \mathcal{F}_n] \to I_A$ with probability 1. Compare Lebesgue's density theorem.

35.16. Theorems 35.6 and 35.9 have analogues in Hilbert space. For $n \le \infty$, let P_n be the perpendicular projection on a subspace M_n. Then $P_n x \to P_\infty x$ for all x if either (a) $M_1 \subset M_2 \subset \cdots$ and M_∞ is the closure of $\bigcup_{n < \infty} M_n$ or (b) $M_1 \supset M_2 \supset \cdots$ and $M_\infty = \bigcap_{n < \infty} M_n$.

35.17. Suppose that θ has an arbitrary distribution, and suppose that, conditionally on θ, the random variables $Y_1, Y_2, \ldots$ are independent and normally distributed with mean θ and variance σ^2. Construct such a sequence $\{\theta, Y_1, Y_2, \ldots\}$. Prove (35.31).

35.18. It is shown on p. 471 that optional stopping has no effect on likelihood ratios This is not true of tests of significance. Suppose that $X_1, X_2, \ldots$ are independent and identically distributed and assume the values 1 and 0 with probabilities p and $1 - p$. Consider the null hypothesis that $p = \frac{1}{2}$ and the alternative that $p > \frac{1}{2}$. The usual .05-level test of significance is to reject the null

hypothesis if

(35.40) $\dfrac{2}{\sqrt{n}}(X_1 + \cdots + X_n - \tfrac{1}{2}n) > 1.645.$

For this test the chance of falsely rejecting the null hypothesis is approximately $P[N > 1.645] \approx .05$ if n is large and fixed. Suppose that n is not fixed in advance of sampling, and show by the law of the iterated logarithm that, even if p is, in fact, $\tfrac{1}{2}$, there are with probability 1 infinitely many n for which (35.40) holds.

35.19. (a) Suppose that (35.32) and (35.33) hold. Suppose further that, for constants s_n^2, $s_n^{-2}\sum_{k=1}^{n}\sigma_k^2 \to_p 1$ and $s_n^{-2}\sum_{k=1}^{n}E[Y_k^2 I_{[|Y_k| \geq s_n]}] \to 0$, and show that $s_n^{-1}\sum_{k=1}^{n}Y_k \Rightarrow N$. *Hint:* Simplify the proof of Theorem 35.11.

(b) *The Lindeberg-Lévy theorem for martingales* Suppose that

$$\ldots, Y_{-1}, Y_0, Y_1, \ldots$$

is stationary and ergodic (p. 494) and that

$$E[Y_k^2] < \infty \quad \text{and} \quad E[Y_k \| Y_{k-1}, Y_{k-2}, \ldots] = 0.$$

Prove that $\sum_{k=1}^{n}Y_k /\sqrt{n}$ is asymptotically normal. *Hint:* Use Theorem 36.4 and the remark following the statement of Lindeberg's Theorem 27.2.

35.20. 24.4↑ Suppose that the σ-field $\mathscr{F}_\infty$ in Problem 24.4 is trivial. Deduce from Theorem 35.9 that $P[A\|T^{-n}\mathscr{F}] \to P[A\|\mathscr{F}_\infty] = P(A)$ with probability 1, and conclude that T is mixing.

CHAPTER 7

Stochastic Processes

SECTION 36. KOLMOGOROV'S EXISTENCE THEOREM

Stochastic Processes

A *stochastic process* is a collection $[X_t: t \in T]$ of random variables on a probability space $(\Omega, \mathscr{F}, P)$. The sequence of gambler's fortunes in Section 7, the sequences of independent random variables in Section 22, the martingales in Section 35—all these are stochastic processes for which $T = \{1, 2, \dots\}$. For the Poisson process $[N_t: t \geq 0]$ of Section 23, $T = [0, \infty)$. For all these processes the points of T are thought of as representing *time*. In most cases, T is the set of integers and time is *discrete*, or else T is an interval of the line and time is *continuous*. For the general theory of this section, however, T can be quite arbitrary.

Finite-Dimensional Distributions

A process is usually described in terms of distributions it induces in Euclidean spaces. For each k-tuple $(t_1, \dots, t_k)$ of distinct elements of T, the random vector $(X_{t_1}, \dots, X_{t_k})$ has over R^k some distribution $\mu_{t_1 \dots t_k}$:

$$(36.1) \qquad \mu_{t_1 \dots t_k}(H) = P\left[(X_{t_1}, \dots, X_{t_k}) \in H \right], \qquad H \in \mathscr{R}^k.$$

These probability measures $\mu_{t_1 \dots t_k}$ are the *finite-dimensional distributions* of the stochastic process $[X_t: t \in T]$. The system of finite-dimensional distributions does not completely determine the properties of the process. For example, the Poisson process $[N_t: t \geq 0]$ as defined by (23.5) has sample paths (functions $N_t(\omega)$ with ω fixed and t varying) that are step functions. But (23.28) defines a process that has the same finite-dimensional distributions and has sample paths that are *not* step functions. Nevertheless, the first step in a general theory is to construct processes for given systems of finite-dimensional distributions.

482

Now (36.1) implies two consistency properties of the system $\mu_{i_1 \ldots i_k}$. Suppose the H in (36.1) has the form $H = H_1 \times \cdots \times H_k$ ($H_i \in \mathscr{R}^1$), and consider a permutation π of $(1, 2, \ldots, k)$. Since $[(X_{i_1}, \ldots, X_{i_k}) \in (H_1 \times \cdots \times H_k)]$ and $[(X_{i_{\pi 1}}, \ldots, X_{i_{\pi k}}) \in (H_{\pi 1} \times \cdots \times H_{\pi k})]$ are the same event, it follows by (36.1) that

$$(36.2) \qquad \mu_{i_1 \ldots i_k}(H_1 \times \cdots \times H_k) = \mu_{i_{\pi 1} \ldots i_{\pi k}}(H_{\pi 1} \times \cdots \times H_{\pi k}).$$

For example, if $\mu_{s,t} = \nu \times \nu'$, then necessarily $\mu_{t,s} = \nu' \times \nu$.

The second consistency condition is

$$(36.3) \qquad \mu_{i_1 \ldots i_{k-1}}(H_1 \times \cdots \times H_{k-1}) = \mu_{i_1 \ldots i_{k-1} i_k}(H_1 \times \cdots \times H_{k-1} \times R^1).$$

This is clear because $(X_{i_1}, \ldots, X_{i_{k-1}})$ lies in $H_1 \times \cdots \times H_{k-1}$ if and only if $(X_{i_1}, \ldots, X_{i_{k-1}}, X_{i_k})$ lies in $H_1 \times \cdots \times H_{k-1} \times R^1$.

Measures $\mu_{i_1 \ldots i_k}$ coming from a process $[X_t: t \in T]$ via (36.1) necessarily satisfy (36.2) and (36.3). *Kolmogorov's existence theorem* says conversely that if a given system of measures satisfies the two consistency conditions, then there exists a stochastic process having these finite-dimensional distributions. The proof is a construction, one which is more easily understood if (36.2) and (36.3) are combined into a single condition.

Define $\varphi_\pi: R^k \to R^k$ by

$$\varphi_\pi(x_1, \ldots, x_k) = (x_{\pi^{-1} 1}, \ldots, x_{\pi^{-1} k});$$

φ_π applies the permutation π to the coordinates (for example, if π sends x_3 to first position, then $\pi^{-1} 1 = 3$). Since $\varphi_\pi^{-1}(H_1 \times \cdots \times H_k) = H_{\pi 1} \times \cdots \times H_{\pi k}$, it follows from (36.2) that

$$\mu_{i_{\pi 1} \ldots i_{\pi k}} \varphi_\pi^{-1}(H) = \mu_{i_1 \ldots i_k}(H)$$

for rectangles H. But then

$$(36.4) \qquad \mu_{i_1 \ldots i_k} = \mu_{i_{\pi 1} \ldots i_{\pi k}} \varphi_\pi^{-1}.$$

Similarly, if $\varphi: R^k \to R^{k-1}$ is the projection $\varphi(x_1, \ldots, x_k) = (x_1, \ldots, x_{k-1})$, then (36.3) is the same thing as

$$(36.5) \qquad \mu_{i_1 \ldots i_{k-1}} = \mu_{i_1 \ldots i_k} \varphi^{-1}.$$

The conditions (36.4) and (36.5) have a common extension. Suppose that $(u_1, \ldots, u_m)$ is an m-tuple of distinct elements of T and that each element of $(t_1, \ldots, t_k)$ is also an element of $(u_1, \ldots, u_m)$. Then $(t_1, \ldots, t_k)$ must be the initial segment of some permutation of $(u_1, \ldots, u_m)$; that is, $k \leq m$ and there

is a permutation π of $(1, 2, \ldots, m)$ such that

$$(u_{\pi^{-1}1}, \ldots, u_{\pi^{-1}m}) = (t_1, \ldots, t_k, t_{k+1}, \ldots, t_m),$$

where $t_{k+1}, \ldots, t_m$ are elements of $(u_1, \ldots, u_m)$ that do not appear in $(t_1, \ldots, t_k)$. Define $\psi: R^m \to R^k$ by

$$(36.6) \qquad \psi(x_1, \ldots, x_m) = (x_{\pi^{-1}1}, \ldots, x_{\pi^{-1}k});$$

ψ applies π to the coordinates and then projects onto the first k of them. Since $\psi(X_{u_1}, \ldots, X_{u_m}) = (X_{t_1}, \ldots, X_{t_k})$,

$$(36.7) \qquad \mu_{t_1 \cdots t_k} = \mu_{u_1 \cdots u_m} \psi^{-1}.$$

This contains (36.4) and (36.5) as special cases, but as ψ is a coordinate permutation followed by a sequence of projections of the form $(x_1, \ldots, x_l) \to (x_1, \ldots, x_{l-1})$, it is also a consequence of these special cases.

Product Spaces

The standard construction of the general process involves product spaces. Let T be an arbitrary index set, and let R^T be the collection of all real functions on T—all maps from T into the real line. If $T = \{1, 2, \ldots, k\}$, a real function on T can be identified with a k-tuple $(x_1, \ldots, x_k)$ of real numbers, and so R^T can be identified with k-dimensional Euclidean space R^k. If $T = \{1, 2, \ldots\}$, a real function on T is a sequence $\{x_1, x_2, \ldots\}$ of real numbers. If T is an interval, R^T consists of all real functions, however irregular, on the interval. The theory of R^T is an elaboration of the theory of the analogous but simpler space S^∞ of Section 2 (p. 27).

Whatever the set T may be, an element of R^T will be denoted x. The value of x at t will be denoted $x(t)$ or x_t, depending on whether x is viewed as a function of t with domain T or as a vector with components indexed by the elements t of T. Just as R^k can be regarded as the Cartesian product of k copies of the real line, R^T can be regarded as a *product space*—a product of copies of the real line, one copy for each t in T.

For each t define a mapping $Z_t: R^T \to R^1$ by

$$(36.8) \qquad Z_t(x) = x(t) = x_t.$$

The Z_t are called the *coordinate functions* or *projections*. When later on a probability measure has been defined on R^T, the Z_t will be random variables, the *coordinate variables*. Frequently, the value $Z_t(x)$ is instead denoted $Z(t, x)$. If x is fixed, $Z(\cdot, x)$ is a real function on T and is, in fact, nothing other than $x(\cdot)$—that is, x itself. If t is fixed, $Z(t, \cdot)$ is a real function on R^T and is identical with the function Z_t defined by (36.8).

There is a natural generalization to R^T of the idea of the σ-field of k-dimensional Borel sets. Let $\mathscr{R}^T$ be the σ-field generated by all the coordinate functions Z_t, $t \in T$: $\mathscr{R}^T = \sigma[Z_t: t \in T]$. It is generated by the sets of the form

$$\left[x \in R^T: Z_t(x) \in H \right] = \left[x \in R^T: x_t \in H \right]$$

for $t \in T$ and $H \in \mathscr{R}^1$. If $T = \{1, 2, \ldots, k\}$, then $\mathscr{R}^T$ coincides with $\mathscr{R}^k$.

Consider the class $\mathscr{R}_0^T$ consisting of the sets of the form

(36.9)
$$A = \left[x \in R^T: \left(Z_{t_1}(x), \ldots, Z_{t_k}(x) \right) \in H \right]$$

$$= \left[x \in R^T: \left(x_{t_1}, \ldots, x_{t_k} \right) \in H \right],$$

where k is an integer, $(t_1, \ldots, t_k)$ is a k-tuple of distinct points of T, and $H \in \mathscr{R}^k$. Sets of this form, elements of $\mathscr{R}_0^T$, are called *finite-dimensional sets*, or *cylinders*. Of course, $\mathscr{R}_0^T$ generates $\mathscr{R}^T$. Now $\mathscr{R}_0^T$ is not a σ-field, does not coincide with $\mathscr{R}^T$ (unless T is finite), but the following argument shows that it is a field.
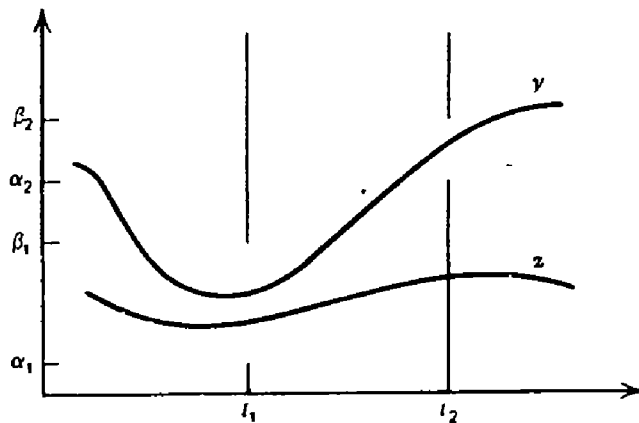
If T is an interval, the cylinder $[x \in R^T: \alpha_1 < x(t_1) \le \beta_1, \alpha_2 < x(t_2) \le \beta_2]$ consists of the functions that go through the two gates shown; y lies in the cylinder and z does not (they need not be continuous functions, of course)

The complement of (36.9) is $R^T - A = [x \in R^T: (x_{t_1}, \ldots, x_{t_k}) \in R^k - H]$, and so $\mathscr{R}_0^T$ is closed under complementation. Suppose that A is given by (36.9) and B is given by

(36.10)
$$B = \left[x \in R^T: \left(x_{s_1}, \ldots, x_{s_j} \right) \in I \right],$$

where $I \in \mathscr{R}^j$. Let $(u_1, \ldots, u_m)$ be an m-tuple containing all the t_α and all the s_β. Now $(t_1, \ldots, t_k)$ must be the initial segment of some permutation of $(u_1, \ldots, u_m)$, and if ψ is as in (36.6) and $H' = \psi^{-1}H$, then $H' \in \mathscr{R}^m$ and A is

given by

(36.11) $A = \left[x \in R^T : (x_{u_1}, \ldots, x_{u_m}) \in H' \right]$

as well as by (36.9). Similarly, B can be put in the form

(36.12) $B = \left[x \in R^T : (x_{u_1}, \ldots, x_{u_m}) \in I' \right]$,

where $I' \in \mathscr{R}^m$. But then

(36.13) $A \cup B = \left[x \in R^T : (x_{u_1}, \ldots, x_{u_m}) \in H' \cup I' \right]$.

Since $H' \cup I' \in \mathscr{R}^m$, $A \cup B$ is a cylinder. This proves that $\mathscr{R}_0^T$ *is a field such that* $\mathscr{R}^T = \sigma(\mathscr{R}_0^T)$.

The Z_t are measurable functions on the measurable space $(R^T, \mathscr{R}^T)$. If P is a probability measure on $\mathscr{R}^T$, then $[Z_t : t \in T]$ is a stochastic process on $(R^T, \mathscr{R}^T, P)$, the *coordinate-variable process*.

Kolmogorov's Existence Theorem

The existence theorem can be stated two ways:

Theorem 36.1. *If* $\mu_{t_1 \ldots t_k}$ *are a system of distributions satisfying the consistency conditions* (36.2) *and* (36.3), *then there is a probability measure P on $\mathscr{R}^T$ such that the coordinate-variable process* $[Z_t : t \in T]$ *on* $(R^T, \mathscr{R}^T, P)$ *has the* $\mu_{t_1 \ldots t_k}$ *as its finite-dimensional distributions.*

Theorem 36.2. *If* $\mu_{t_1 \ldots t_k}$ *are a system of distributions satisfying the consistency conditions* (36.2) *and* (36.3), *then there exists on some probability space* $(\Omega, \mathscr{F}, P)$ *a stochastic process* $[X_t : t \in T]$ *having the* $\mu_{t_1 \ldots t_k}$ *as its finite-dimensional distributions.*

For many purposes the underlying probability space is irrelevant, the joint distributions of the variables in the process being all that matters, so that the two theorems are equally useful. As a matter of fact, they are equivalent anyway. Obviously, the first implies the second. To prove the converse, suppose that the process $[X_t : t \in T]$ on $(\Omega, \mathscr{F}, P)$ has finite-dimensional distributions $\mu_{t_1 \ldots t_k}$, and define a map $\xi : \Omega \to R^T$ by the requirement

(36.14) $Z_t(\xi(\omega)) = X_t(\omega), \quad t \in T.$

For each ω, $\xi(\omega)$ is an element of R^T, a real function on T, and the

requirement is that $X_t(\omega)$ be its value at t. Clearly,

$$(36.15) \qquad \xi^{-1}\Big[x \in R^T \colon \big(Z_{t_1}(x), \ldots, Z_{t_k}(x)\big) \in H\Big]$$

$$= \Big[\omega \in \Omega \colon \big(Z_{t_1}(\xi(\omega)), \ldots, Z_{t_k}(\xi(\omega))\big) \in H\Big]$$

$$= \Big[\omega \in \Omega \colon \big(X_{t_1}(\omega), \ldots, X_{t_k}(\omega)\big) \in H\Big];$$

since the X_t are random variables, measurable $\mathscr{F}$, this set lies in $\mathscr{F}$ if $H \in \mathscr{R}^k$. Thus $\xi^{-1}A \in \mathscr{F}$ for $A \in \mathscr{R}_0^T$, and so (Theorem 13.1) ξ is measurable $\mathscr{F}/\mathscr{R}^T$. By (36.15) and the assumption that $[X_t \colon t \in T]$ has finite-dimensional distributions $\mu_{t_1 \ldots t_k}$, $P\xi^{-1}$ (see (13.7)) satisfies

$$(36.16) \quad P\xi^{-1}\Big[x \in R^T \colon \big(Z_{t_1}(x), \ldots, Z_{t_k}(x)\big) \in H\Big]$$

$$= P\Big[\omega \in \Omega \colon \big(X_{t_1}(\omega), \ldots, X_{t_k}(\omega)\big) \in H\Big] = \mu_{t_1 \ldots t_k}(H).$$

Thus the coordinate-variable process $[Z_t \colon t \in T]$ on $(R^T, \mathscr{R}^T, P\xi^{-1})$ also has finite-dimensional distributions $\mu_{t_1 \ldots t_k}$.

Therefore, to prove either of the two versions of Kolmogorov's existence theorem is to prove the other one as well.

Example 36.1. Suppose that T is finite, say $T = \{1, 2, \ldots, k\}$. Then $(R^T, \mathscr{R}^T)$ is $(R^k, \mathscr{R}^k)$, and taking $P = \mu_{1,2,\ldots,k}$ satisfies the requirements of Theorem 36.1. ∎

Example 36.2. Suppose that $T = \{1, 2, \ldots\}$ and

$$(36.17) \qquad\qquad \mu_{t_1 \ldots t_k} = \mu_{t_1} \times \cdots \times \mu_{t_k},$$

where $\mu_1, \mu_2, \ldots$ are probability distributions on the line. The consistency conditions are easily checked, and the probability measure P guaranteed by Theorem 36.1 is *product measure* on the product space $(R^T, \mathscr{R}^T)$. But by Theorem 20.4 there exists on some $(\Omega, \mathscr{F}, P)$ an independent sequence $X_1, X_2, \ldots$ of random variables with respective distributions $\mu_1, \mu_2, \ldots$; then (36.17) is the distribution of $(X_{t_1}, \ldots, X_{t_n})$. For the special case (36.17), Theorem 36.2 (and hence Theorem 36.1) was thus proved in Section 20. The existence of independent sequences with prescribed distributions was the measure-theoretic basis of all the probabilistic developments in Chapters 4, 5, and 6: even dependent processes like the Poisson were constructed from independent sequences. The existence of independent sequences can also be made the basis of a proof of Theorems 36.1 and 36.2 in their full generality; see the second proof below. ∎

Example 36.3. The preceding example has an analogue in the space S^∞ of sequences (2.15). Here the finite set S plays the role of R^1, the $z_n(\cdot)$ are analogues of the $Z_n(\cdot)$, and the product measure defined by (2.21) is the analogue of the product measure specified by (36.17) with $\mu_i = \mu$. See also Example 24.2. The theory for S^∞ is simple because S is finite: see Theorem 2.3 and the lemma it depends on. ∎

Example 36.4. If T is a subset of the line, it is convenient to use the order structure of the line and take the $\mu_{s_1 \cdots s_k}$ to be specified initially only for k-tuples $(s_1, \ldots, s_k)$ that are in increasing order:

$$(36.18) \qquad s_1 < s_2 < \cdots < s_k.$$

It is natural for example to specify the finite-dimensional distributions for the Poisson processes for increasing sequences of time points alone; see (23.27).

Assume that the $\mu_{s_1 \cdots s_k}$ for k-tuples satisfying (36.18) have the consistency property

$$(36.19) \quad \mu_{s_1 \cdots s_{i-1} s_{i+1} \cdots s_k}(H_1 \times \cdots \times H_{i-1} \times H_{i+1} \times \cdots \times H_k)$$

$$= \mu_{s_1 \cdots s_k}(H_1 \times \cdots \times H_{i-1} \times R^1 \times H_{i+1} \times \cdots \times H_k).$$

For given $s_1, \ldots, s_k$ satisfying (36.18), take $(X_{s_1}, \ldots, X_{s_k})$ to have distribution $\mu_{s_1 \cdots s_k}$. If $t_1, \ldots, t_k$ is a permutation of $s_1, \ldots, s_k$, take $\mu_{t_1 \cdots t_k}$ to be the distribution of $(X_{t_1}, \ldots, X_{t_k})$:

$$(36.20) \qquad \mu_{t_1 \cdots t_k}(H_1 \times \cdots \times H_k) = P[X_{t_i} \in H_i, i \le k].$$

This unambiguously defines a collection of finite-dimensional distributions. Are they consistent?

If $t_{\pi 1}, \ldots, t_{\pi k}$ is a permutation of $t_1, \ldots, t_k$, then it is also a permutation of $s_1, \ldots, s_k$, and by the definition (36.20), $\mu_{t_{\pi 1} \cdots t_{\pi k}}$ is the distribution of $(X_{t_{\pi 1}}, \ldots, X_{t_{\pi k}})$, which immediately gives (36.2), the first of the consistency conditions. Because of (36.19), $\mu_{s_1 \cdots s_{i-1} s_{i+1} \cdots s_k}$ is the distribution of $(X_{s_1}, \ldots, X_{s_{i-1}}, X_{s_{i+1}}, \ldots, X_{s_k})$, and if $t_k = s_i$, then $t_1, \ldots, t_{k-1}$ is a permutation of $s_1, \ldots, s_{i-1}, s_{i+1}, \ldots, s_k$, which are in increasing order. By the definition (36.20) applied to $t_1, \ldots, t_{k-1}$ it therefore follows that $\mu_{t_1 \cdots t_{k-1}}$ is the distribution of $(X_{t_1}, \ldots, X_{t_{k-1}})$. But this gives (36.3), the second of the consistency conditions.

It will therefore follow from the existence theorem that if $T \subset R^1$ and $\mu_{s_1 \cdots s_k}$ is defined for all k-tuples in increasing order, and if (36.19) holds, then there exists a stochastic process $[X_t: t \in T]$ satisfying (36.1) for increasing $t_1, \ldots, t_k$. ∎

Two proofs of Kolmogorov's existence theorem will be given. The first is based on the extension theorem of Section 3.

FIRST PROOF OF KOLMOGOROV'S THEOREM. Consider the first formulation, Theorem 36.1. If A is the cylinder (36.9), define

$$(36.21) \qquad\qquad P(A) = \mu_{t_1, \ldots t_k}(H).$$

This gives rise to the question of consistency because A will have other representations as a cylinder. Suppose, in fact, that A coincides with the cylinder B defined by (36.10). As observed before, if $(u_1, \ldots, u_m)$ contains all the t_α and s_β, A is also given by (36.11), where $H' = \psi^{-1}H$ and ψ is defined in (36.6). Since the consistency conditions (36.2) and (36.3) imply the more general one (36.7), $P(A) = \mu_{t_1 \ldots t_k}(H) = \mu_{u_1 \ldots u_m}(H')$. Similarly, (36.10) has the form (36.12), and $P(B) = \mu_{s_1 \ldots s_l}(I) = \mu_{u_1 \ldots u_m}(I')$. Since the u_γ are distinct, for any real numbers $z_1, \ldots, z_m$ there are points x of R^T for which $(x_{u_1}, \ldots, x_{u_m}) = (z_1, \ldots, z_m)$. From this it follows that if the cylinders (36.11) and (36.12) coincide, then $H' = I'$. Hence $A = B$ implies that $P(A) = \mu_{u_1 \ldots u_m}(H') = \mu_{u_1 \ldots u_m}(I') = P(B)$, and the definition (36.21) is indeed consistent.

Now consider disjoint cylinders A and B. As usual, the index sets may be taken identical. Assume then that A is given by (36.11) and B by (36.12), so that (36.13) holds. If $H' \cap I'$ were nonempty, then $A \cap B$ would be nonempty as well. Therefore, $H' \cap I' = \varnothing$, and

$$P(A \cup B) = \mu_{u_1 \ldots u_m}(H' \cup I')$$

$$= \mu_{u_1 \ldots u_m}(H') + \mu_{u_1 \ldots u_m}(I') = P(A) + P(B).$$

Therefore, P is finitely additive on $\mathscr{R}_0^T$. Clearly, $P(R^T) = 1$.

Suppose that P is shown to be countably additive on $\mathscr{R}_0^T$. By Theorem 3.1, P will then extend to a probability measure on $\mathscr{R}^T$. By the way P was defined on $\mathscr{R}_0^T$,

$$(36.22) \qquad P\big[x \in R^T \colon \big(Z_{t_1}(x), \ldots, Z_{t_k}(x)\big) \in H\big] = \mu_{t_1 \ldots t_k}(H),$$

and therefore the coordinate process $[Z_t \colon t \in T]$ will have the required finite-dimensional distributions.

It suffices, then, to prove P countably additive on $\mathscr{R}_0^T$, and this will follow if $A_n \in \mathscr{R}_0^T$ and $A_n \downarrow \varnothing$ together imply $P(A_n) \downarrow 0$ (see Example 2.10). Suppose that $A_1 \supset A_2 \supset \cdots$ and that $P(A_n) \geq \epsilon > 0$ for all n. The problem is to show that $\bigcap_n A_n$ must be nonempty. Since $A_n \in \mathscr{R}_0^T$, and since the index set involved in the specification of a cylinder can always be permuted and

expanded, there exists a sequence $t_1, t_2, \ldots$ of points in T for which

$$A_n = \left[x \in R^T: (x_{t_1}, \ldots, x_{t_n}) \in H_n \right],$$

where[†] $H_n \in \mathscr{R}^n$.

Of course, $P(A_n) = \mu_{t_1} \ldots t_n(H_n)$. By Theorem 12.3 (regularity), there exists inside H_n a compact set K_n such that $\mu_{t_1} \ldots t_n(H_n - K_n) < \epsilon/2^{n+1}$. If $B_n = [x \in R^T: (x_{t_1}, \ldots, x_{t_n}) \in K_n]$, then $P(A_n - B_n) < \epsilon/2^{n+1}$. Put $C_n = \bigcap_{k=1}^{n} B_k$. Then $C_n \subset B_n \subset A_n$ and $P(A_n - C_n) < \epsilon/2$, so that $P(C_n) > \epsilon/2 > 0$. Therefore, $C_n \subset C_{n-1}$ and C_n is nonempty.

Choose a point $x^{(n)}$ of R^T in C_n. If $n \geq k$, then $x^{(n)} \in C_n \subset C_k \subset B_k$ and hence $(x_{t_1}^{(n)}, \ldots, x_{t_k}^{(n)}) \in K_k$. Since K_k is bounded, the sequence $\{x_{t_k}^{(1)}, x_{t_k}^{(2)}, \ldots\}$ is bounded for each k. By the diagonal method [A14] select an increasing sequence $n_1, n_2, \ldots$ of integers such that $\lim_i x_{t_k}^{(n_i)}$ exists for each k. There is in R^T some point x whose t_kth coordinate is this limit for each k. But then, for each k, $(x_{t_1}, \ldots, x_{t_k})$ is the limit as $i \to \infty$ of $(x_{t_1}^{(n_i)}, \ldots, x_{t_k}^{(n_i)})$ and hence lies in K_k. But that means that x itself lies in B_k and hence in A_k. Thus $x \in \bigcap_{k=1}^{\infty} A_k$, which completes the proof.[‡]   ■

The second proof of Kolmogorov's theorem goes in two stages, first for countable T, then for general T.[*]

SECOND PROOF FOR COUNTABLE T. The result for countable T will be proved in its second formulation, Theorem 36.2. It is no restriction to enumerate T as $\{t_1, t_2, \ldots\}$ and then to identify t_n with n; in other words, it is no restriction to assume that $T = \{1, 2, \ldots\}$. Write μ_n in place of $\mu_{1,2,\ldots,n}$.

By Theorem 20.4 there exists on a probability space $(\Omega, \mathscr{F}, P)$ (which can be taken to be the unit interval) an independent sequence $U_1, U_2, \ldots$ of random variables each uniformly distributed over $(0,1)$. Let F_1 be the distribution function corresponding to μ_1. If the "inverse" g_1 of F_1 is defined over $(0,1)$ by $g_1(s) = \inf[x: s \leq F_1(x)]$, then $X_1 = g_1(U_1)$ has distribution μ_1 by the usual argument: $P[g_1(U_1) \leq x] = P[U_1 \leq F_1(x)] = F_1(x)$.

The problem is to construct $X_2, X_3, \ldots$ inductively in such a way that

$$(36.23) \qquad\qquad X_k = h_k(U_1, \ldots, U_k)$$

for a Borel function h_k and $(X_1, \ldots, X_n)$ has the distribution μ_n. Assume that $X_1, \ldots, X_{n-1}$ have been defined $(n \geq 2)$: they have joint distribution μ_{n-1} and (36.23) holds for $k \leq n-1$. The idea now is to construct an appropriate conditional distribution function $F_n(x|x_1, \ldots, x_{n-1})$; here $F_n(x|X_1(\omega), \ldots, X_{n-1}(\omega))$ will have the value $P[X_n \leq x \| X_1, \ldots, X_{n-1}]_\omega$ would have if X_n were already defined. If $g_n(\cdot | x_1, \ldots, x_{n-1})$

[†]In general, A_n will involve indices $t_1, \ldots, t_{a_n}$, where $a_1 < a_2 < \cdots$. For notational simplicity a_n is taken as n. As a matter of fact, this can be arranged anyway: Take $A'_{a_n} = A_n$, $A'_k = [x: (x_{t_1}, \ldots, x_{t_k}) \in R^k] = R^T$ for $k < a_1$, and $A'_k = [x: (x_{t_1}, \ldots, x_{t_k}) \in H_n \times R^{k-a_n}] = A_n$ for $a_n < k \leq a_{n+1}$. Now relabel A'_n as A_n.

[‡]The last part of the argument is, in effect, the proof that a countable product of compact sets is compact.

[*]This second proof, which may be omitted, uses the conditional-probability theory of Section 33.

is the "inverse" function, then $X_n(\omega) = g_n(U_n(\omega)|X_1(\omega), \ldots, X_{n-1}(\omega))$ will by the usual argument have the right conditional distribution given $X_1, \ldots, X_{n-1}$, so that $(X_1, \ldots, X_{n-1}, X_n)$ will have the right distribution over R^n.

To construct the conditional distribution function, apply Theorem 33.3 in $(R^n, \mathscr{R}^n, \mu_n)$ to get a conditional distribution of the last coordinate of $(x_1, \ldots, x_n)$ given the first $n - 1$ of them. This will have (Theorem 20.1) the form $\nu(H; x_1, \ldots, x_{n-1})$; it is a probability measure as H varies over $\mathscr{R}^1$, and

$$\int_{(x_1, \ldots, x_{n-1}) \in M} \nu(H; x_1, \ldots, x_{n-1}) \, d\mu_n(x_1, \ldots, x_n)$$

$$= \mu_n[x \in R^n \cdot (x_1, \ldots, x_{n-1}) \in M, x_n \in H].$$

Since the integrand involves only $x_1, \ldots, x_{n-1}$, and since μ_n by consistency projects to μ_{n-1} under the map $(x_1, \ldots, x_n) \to (x_1, \ldots, x_{n-1})$, a change of variable gives

$$\int_M \nu(H; x_1, \ldots, x_{n-1}) \, d\mu_{n-1}(x_1, \ldots, x_{n-1})$$

$$= \mu_n[x \in R^n : (x_1, \ldots, x_{n-1}) \in M, x_n \in H].$$

Define $F_n(x|x_1, \ldots, x_{n-1}) = \nu((-\infty, x]; x_1, \ldots, x_{n-1})$. Then $F_n(\cdot|x_1, \ldots, x_{n-1})$ is a probability distribution function over the line, $F_n(x|\cdot)$ is a Borel function over R^{n-1}, and

$$\int_M F_n(x|x_1, \ldots, x_{n-1}) \, d\mu_{n-1}(x_1, \ldots, x_{n-1})$$

$$= \mu_n[x \in R^n : (x_1, \ldots, x_{n-1}) \in M, x_n \leq x].$$

Put $g_n(u|x_1, \ldots, x_{n-1}) = \inf[x : u \leq F_n(x|x_1, \ldots, x_{n-1})]$ for $0 < u < 1$. Since $F_n(x|x_1, \ldots, x_{n-1})$ is nondecreasing and right-continuous in x, $g_n(u|x_1, \ldots, x_{n-1}) \leq x$ if and only if $u \leq F_n(x|x_1, \ldots, x_{n-1})$. Set $X_n = g_n(U_n|X_1, \ldots, X_{n-1})$. Since $(X_1, \ldots, X_{n-1})$ has distribution μ_{n-1} and by (36.23) is independent of U_n, an application of (20.30) gives

$$P[(X_1, \ldots, X_{n-1}) \in M, X_n \leq x]$$

$$= P[(X_1, \ldots, X_{n-1}) \in M, U_n \leq F_n(x|X_1, \ldots, X_{n-1})]$$

$$= \int_M P[U_n \leq F_n(x|x_1, \ldots, x_{n-1})] \, d\mu_{n-1}(x_1, \ldots, x_{n-1})$$

$$= \int_M F_n(x|x_1, \ldots, x_{n-1}) \, d\mu_{n-1}(x_1, \ldots, x_{n-1})$$

$$= \mu_n[x \in R^n : (x_1, \ldots, x_{n-1}) \in M, x_n \leq x].$$

Thus $(X_1, \ldots, X_n)$ has distribution μ_n. Note that X_n, as a function of $X_1, \ldots, X_{n-1}$ and U_n, is a function of $U_1, \ldots, U_n$ because (36.23) was assumed to hold for $k < n$. Hence (36.23) holds for $k = n$ as well. ∎

SECOND PROOF FOR GENERAL T. Consider $(R^T, \mathscr{R}^T)$ once again. If $S \subset T$, let $\mathscr{F}_S = \sigma[Z_t : t \in S]$. Then $\mathscr{F}_S \subset \mathscr{F}_T = \mathscr{R}^T$.

Suppose that S is countable. By the case just treated, there exists a process $[X_t : t \in S]$ on some $(\Omega, \mathscr{F}, P)$—the space and the process depend on S—such that $(X_{t_1}, \ldots, X_{t_k})$ has distribution $\mu_{t_1 \ldots t_k}$ for every k-tuple $(t_1, \ldots, t_k)$ from S. Define a map $\xi : \Omega \to R^T$ by requiring that

$$Z_t(\xi(\omega)) = \begin{cases} X_t(\omega) & \text{if } t \in S, \\ 0 & \text{if } t \notin S. \end{cases}$$

Now (36.15) holds as before if $t_1, \ldots, t_k$ all lie in S, and so ξ is measurable $\mathscr{F}/\mathscr{F}_S$. Further, (36.16) holds for $t_1, \ldots, t_k$ in S. Put $P_S = P\xi^{-1}$ on $\mathscr{F}_S$. Then P_S is a probability measure on $(R^T, \mathscr{F}_S)$, and

$$(36.24) \qquad P_S\Big[x \in R^T : \big(Z_{t_1}(x), \ldots, Z_{t_k}(x)\big) \in H\Big] = \mu_{t_1 \ldots t_k}(H)$$

if $H \in \mathscr{R}^k$ and $t_1, \ldots, t_k$ all lie in S. (The various spaces $(\Omega, \mathscr{F}, P)$ and processes $[X_t : t \in S]$ now become irrelevant.)

If $S_0 \subset S$, and if A is a cylinder (36.9) for which the $t_1, \ldots, t_k$ lie in S_0, then $P_{S_0}(A)$ and $P_S(A)$ coincide, their common value being $\mu_{t_1 \ldots t_k}(H)$. Since these cylinders generate $\mathscr{F}_{S_0}$, $P_{S_0}(A) = P_S(A)$ for all A in $\mathscr{F}_{S_0}$. If A lies both in $\mathscr{F}_{S_1}$ and $\mathscr{F}_{S_2}$, then $P_{S_1}(A) = P_{S_1 \cup S_2}(A) = P_{S_2}(A)$. Thus $P(A) = P_S(A)$ consistently defines a set function on the class $\bigcup_S \mathscr{F}_S$, the union extending over the countable subsets S of T. If A_n lies in this union and $A_n \in \mathscr{F}_{S_n}$ (S_n countable), then $S = \bigcup_n S_n$ is countable and $\bigcup_n A_n$ lies in $\mathscr{F}_S$. Thus $\bigcup_S \mathscr{F}_S$ is a σ-field and so must coincide with $\mathscr{R}^T$. Therefore, P is a probability measure on $\mathscr{R}^T$, and by (36.24) the coordinate process has under P the required finite-dimensional distributions. ∎

The Inadequacy of $\mathscr{R}^T$

Theorem 36.3. *Let $[X_t : t \in T]$ be a family of real functions on Ω.*

(i) *If $A \in \sigma[X_t : t \in T]$ and $\omega \in A$, and if $X_t(\omega) = X_t(\omega')$ for all $t \in T$, then $\omega' \in A$.*

(ii) *If $A \in \sigma[X_t : t \in T]$, then $A \in \sigma[X_t : t \in S]$ for some countable subset S of T.*

PROOF. Define $\xi : \Omega \to R^T$ by $Z_t(\xi(\omega)) = X_t(\omega)$. Let $\mathscr{F} = \sigma[X_t : t \in T]$. By (36.15), ξ is measurable $\mathscr{F}/\mathscr{R}^T$ and hence $\mathscr{F}$ contains the class $[\xi^{-1}M : M \in \mathscr{R}^T]$. The latter class is a σ-field, however, and by (36.15) it contains the sets $[\omega \in \Omega : (X_{t_1}(\omega), \ldots, X_{t_k}(\omega)) \in H]$, $H \in \mathscr{R}^k$, and hence contains the σ-field $\mathscr{F}$ they generate. Therefore

$$(36.25) \qquad \sigma[X_t : t \in T] = \big[\xi^{-1}M : M \in \mathscr{R}^T\big].$$

This is an infinite-dimensional analogue of Theorem 20.1(i).

As for (i), the hypotheses imply that $\omega \in A = \xi^{-1}M$ and $\xi(\omega) = \xi(\omega')$, so that $\omega' \in A$ certainly follows.

For $S \subset T$, let $\mathscr{F}_S = \sigma[X_t: t \in S]$; (ii) says that $\mathscr{F} = \mathscr{F}_T$ coincides with $\mathscr{G} = \bigcup_S \mathscr{F}_S$, the union extending over the countable subsets S of T. If $A_1, A_2, \ldots$ lie in $\mathscr{G}$, A_n lies in $\mathscr{F}_{S_n}$ for some countable S_n, and so $\bigcup_n A_n$ lies in $\mathscr{G}$ because it lies in $\mathscr{F}_S$ for $S = \bigcup_n S_n$. Thus $\mathscr{G}$ is a σ-field, and since it contains the sets $[X_t \in H]$, it contains the σ-field $\mathscr{F}$ they generate. (This part of the argument was used in the second proof of the existence theorem.) ∎

From this theorem it follows that various important sets lie outside the class $\mathscr{R}^T$. Suppose that $T = [0, \infty)$. Of obvious interest is the subset C of R^T consisting of the functions continuous over $[0, \infty)$. But C is not in $\mathscr{R}^T$. For suppose it were. By part (ii) of the theorem (let $\Omega = R^T$ and put $[Z_t: t \in T]$ in the role of $[X_t: t \in T]$), C would lie in $\sigma[Z_t: t \in S]$ for some countable $S \subset [0, \infty)$. But then by part (i) of the theorem (let $\Omega = R^T$ and put $[Z_t: t \in S]$ in the role of $[X_t: t \in T]$), if $x \in C$ and $Z_t(x) = Z_t(y)$ for all $t \in S$, then $y \in C$. From the assumption that C lies in $\mathscr{R}^T$ thus follows the existence of a countable set S such that, if $x \in C$ and $x(t) = y(t)$ for all t in S, then $y \in C$. But whatever countable set S may be, for every continuous x there obviously exist functions y that have discontinuities but agree with x on S. Therefore, cannot lie in $\mathscr{R}^T$.

What the argument shows is this: A set A in R^T cannot lie in $\mathscr{R}^T$ unless there exists a countable subset S of T with the property that, if $x \in A$ and $x(t) = y(t)$ for all t in S, then $y \in A$. Thus A cannot lie in $\mathscr{R}^T$ if it effectively involves all the points t in the sense that, for each x in A and each t in T, it is possible to move x out of A by changing its value at t alone. And C is such a set. For another, consider the set of functions x over $T = [0, \infty)$ that are nondecreasing and assume as values $x(t)$ only nonnegative integers:

$$(36.26) \quad \left[x \in R^{[0,\infty)}: x(s) \le x(t), \, x \le t; \, x(t) \in \{0, 1, \ldots\}, \, t \ge 0 \right].$$

This, too, lies outside $\mathscr{R}^T$.

In Section 23 the Poisson process was defined as follows: Let $X_1, X_2, \ldots$ be independent and identically distributed with the exponential distribution (the probability space Ω on which they are defined may by Theorem 20.4 be taken to be the unit interval with Lebesgue measure). Put $S_0 = 0$ and $S_n = X_1 + \cdots + X_n$. If $S_n(\omega) < S_{n+1}(\omega)$ for $n \ge 0$ and $S_n(\omega) \to \infty$, put $N(t, \omega) = N_t(\omega) = \max[n: S_n(\omega) \le t]$ for $t \ge 0$; otherwise, put $N(t, \omega) = N_t(\omega) = 0$ for $t \ge 0$. Then the stochastic process $[N_t: t \ge 0]$ has the finite-dimensional distributions described by the equations (23.27). The function $N(\cdot, \omega)$ is the *path function* or *sample function*[†] corresponding to ω, and by the construction every path function lies in the set (36.26). This is a good thing if the

---

[†]Other terms are *realization* of the process and *trajectory*.

process is to be a model for, say, calls arriving at a telephone exchange: The sample path represents the history of the calls, its value at t being the number of arrivals up to time t, and so it ought to be nondecreasing and integer-valued.

According to Theorem 36.1, there exists a measure P on R^T for $T = [0, \infty)$ such that the coordinate process $[Z_t: t \geq 0]$ on $(R^T, \mathscr{R}^T, P)$ has the finite-dimensional distributions of the Poisson process. This time does the path function $Z(\cdot, x)$ lie in the set (36.26) with probability 1? Since $Z(\cdot, x)$ is just x itself, the question is whether the set (36.26) has P-measure 1. But this set does not lie in $\mathscr{R}^T$, and so it has no measure at all.

An application of Kolmogorov's existence theorem will always yield a stochastic process with prescribed finite-dimensional distributions, but the process may lack certain path-function properties that it is reasonable to require of it as a model for some natural phenomenon. The special construction of Section 23 gets around this difficulty for the Poisson process, and in the next section a special construction will yield a model for Brownian motion with continuous paths. Section 38 treats a general method for producing stochastic processes that have prescribed finite-dimensional distributions and at the same time have path functions with desirable regularity properties.

A Return to Ergodic Theory*

Write $R^\infty, \mathscr{R}_0^\infty, \mathscr{R}^\infty$ for $R^T, \mathscr{R}_0^T, \mathscr{R}^T$ in the case where the index set $\{0, \pm 1, \pm 2, \dots\}$ consists of all the integers. Then R^∞ is analogous to S^∞ (Sections 2 and 24), except that here the sequences are doubly infinite:

$$x = (\dots, Z_{-1}(x), Z_0(x), Z_1(x), \dots).$$

Let T (not an index set) denote the *shift*: $Z_k(Tx) = Z_{k+1}(x)$, $k = 0, \pm 1, \dots$. This is like the shift in Section 24. Since $A \in \mathscr{R}_0^\infty$ implies $T^{-1}A \in \mathscr{R}_0^\infty$, T is measurable $\mathscr{R}^\infty / \mathscr{R}^\infty$. Clearly, it is invertible.

For a stochastic process $X = (\dots, X_{-1}, X_0, X_1, \dots)$ on $(\Omega, \mathscr{F}, P)$, define $\xi: \Omega \to R^\infty$ by (36.14): $\xi\omega = X(\omega) = (\dots, X_{-1}(\omega), X_0(\omega), X_1(\omega), \dots)$. The measure $P\xi^{-1} = PX^{-1}$ on $(R^\infty, \mathscr{R}^\infty)$ can be viewed as the *distribution* of X. Suppose that X is *stationary* in the sense that, for each $k \geq 1$ and $H \in \mathscr{R}^k$, $P[(X_{n+1}, \dots, X_{n+k}) \in H]$ is the same for all $n = 0, \pm 1, \dots$. Then the shift preserves $P\xi^{-1}$ (use (36.16) and Lemma 1, p. 311). The process X is defined to be *ergodic* if under $P\xi^{-1}$ the shift is ergodic in the sense of Section 24.

In the ergodic case, it follows by the ergodic theorem that

$$(36.27) \qquad \frac{1}{n} \sum_{k=1}^{n} f(T^k x) \to \int_{R^\infty} f(x) P\xi^{-1}(dx)$$

---

*This topic, which requires Section 24, may be omitted.

on a set of $P\xi^{-1}$-measure 1, provided f is measurable $\mathscr{R}^{\infty}$ and integrable. Carry (36.27) back to $(\Omega, \mathscr{F}, P)$ by the inverse set mapping ξ^{-1}. Then

$$(36.28) \quad \frac{1}{n} \sum_{k=1}^{n} f(\dots, X_{k-1}, X_k, X_{k+1}, \dots) \to E[f(\dots, X_{-1}, X_0, X_1, \dots)]$$

with probability 1: (36.28) holds at ω if and only if (36.27) holds at $x = \zeta\omega = X(\omega)$. It is understood that on the left in (36.28), X_k is the center coordinate (the 0th coordinate) of the argument of f, and on the right, X_0 is the center coordinate: *For stationary, ergodic X and integrable f*, (36.28) *holds with probability* 1.

If the X_k are independent, then the Z_k are independent under $P\xi^{-1}$. In this case, $\lim_n P\xi^{-1}(A \cap T^{-n}B) = P\xi^{-1}(A)P\xi^{-1}(B)$ for A and B in $\mathscr{R}_0^{\infty}$, because for large enough n the cylinders A and $T^{-n}B$ depend on disjoint sets of time indices and hence are independent. But then it follows by approximation (Corollary 1 to Theorem 11.4) that the same limit holds for all A and B in $\mathscr{R}^{\infty}$. But for invariant B, this implies $P\xi^{-1}(B^c \cap B) = P\xi^{-1}(B^c)P\xi^{-1}(B)$, so that $P\xi^{-1}(B)$ is 0 or 1, and the shift is ergodic under $P\xi^{-1}$: *If X is stationary and independent, then it is ergodic*.

If f depends on just one coordinate of x, then (36.28) is in the independent case a consequence of the strong law of large numbers, Theorem 22.1. But (36.28) follows by the ergodic theorem even if f involves all the coordinates in some complicated way. Consider now a measurable real function ϕ on R^{∞}. Define $\psi: R^{\infty} \to R^{\infty}$ by

$$\psi(x) = (\dots, \phi(T^{-1}x), \phi(x), \phi(Tx), \dots);$$

here $\phi(x)$ is the center coordinate: $Z_k(\psi(x)) = \phi(T^k x)$. It is easy to show that ψ is measurable $\mathscr{R}^{\infty}/\mathscr{R}^{\infty}$ and commutes with the shift in the sense of Example 24.6. Therefore, T preserves $P\xi^{-1}\psi^{-1}$ if it preserves $P\xi^{-1}$, and it is ergodic under $P\xi^{-1}\psi^{-1}$ if it is ergodic under $P\xi^{-1}$.

This translates immediately into a result on stochastic processes. Define $Y = (\dots, Y_{-1}, Y_0, Y_1, \dots)$ in terms of X by

$$(36.29) \quad Y_n = \phi(\dots, X_{n-1}, X_n, X_{n+1}, \dots),$$

that is to say, $Y(\omega) = \psi(X(\omega)) = \psi\xi\omega$. Since $P\xi^{-1}$ is the distribution of X, $P\xi^{-1}\psi^{-1} = P(\psi\xi)^{-1} = PY^{-1}$ is the distribution of Y:

Theorem 36.4. *If X is stationary and ergodic, in particular if the X_n are independent and identically distributed, then Y as defined by (36.29) is stationary and ergodic.*

This theorem fails if Y is not defined in terms of X in a time-invariant way—if the ϕ in (36.29) is not the same for all n: If $\phi_n(x) = Z_{-n}(x)$ and ϕ is replaced by ϕ_n in (36.29), then $Y_n \equiv X_0$; in this case Y happens to be stationary, but it is not ergodic if the distribution of X_0 does not concentrate at a single point.

Example 36.5. *The autoregressive model.* Let $\phi(x) = \sum_{k=0}^{\infty} \beta^k Z_{-k}(x)$ on the set where the series converges, and take $\phi(x) = 0$ elsewhere. Suppose that $|\beta| < 1$ and that the X_n are independent and identically distributed with finite second moments. Then by Theorem 22.6, $Y_n = \sum_{k=0}^{\infty} \beta^k X_{n-k}$ converges with probability 1, and by Theorem 36.4, the process Y is ergodic. Note that $Y_{n+1} = \beta Y_n + X_{n+1}$ and that X_{n+1} is independent of Y_n. This is the linear autoregressive model of order 1. ∎

The Hewitt–Savage Theorem*

Change notation: Let $(R^\infty, \mathscr{R}^\infty)$ be the product space with $\{1,2,\ldots\}$ as the index set, the space of one-sided sequences. Let P be a probability measure on $\mathscr{R}^\infty$. If the coordinate variables Z_n are independent under P, then by Theorem 22.3, $P(A)$ is 0 or 1 for each A in the tail σ-field $\mathscr{T}$. If the Z_n are also identically distributed under P, a stronger result holds.

Let $\mathscr{S}_n$ be the class of $\mathscr{R}^\infty$-sets A that are invariant under permutations of the first n coordinates: if π is a permutation of $\{1,\ldots,n\}$, then x lies in A if and only if $(Z_{\pi 1}(x),\ldots,Z_{\pi n}(x),Z_{n+1}(x),\ldots)$ does. Then $\mathscr{S}_n$ is a σ-field. Let $\mathscr{S} = \bigcap_{n=1}^\infty \mathscr{S}_n$ be the σ-field of $\mathscr{R}^\infty$-sets invariant under all finite permutations of coordinates. Then $\mathscr{S}$ is larger than $\mathscr{T}$, since, for example, the x-set where $\sum_{k=1}^n Z_k(x) > c_n$ infinitely often lies in $\mathscr{S}$ but not in $\mathscr{T}$.

The *Hewitt–Savage theorem* is a zero–one law for $\mathscr{S}$ in the independent, identically distributed case.

Theorem 36.5. *If the Z_n are independent and identically distributed under P, then $P(A)$ is 0 or 1 for each A in $\mathscr{S}$.*

PROOF. By Corollary 1 to Theorem 11.4, there are for given A and ϵ an n and a set $U = [(Z_1,\ldots,Z_n) \in H]$ $(H \in \mathscr{R}^n)$ such that $P(A \vartriangle U) < \epsilon$. Let $V = [(Z_{n+1},\ldots,Z_{2n}) \in H]$. If the Z_k are independent and identically distributed, then $P(A \vartriangle U)$ is the same as

$$P\big([(Z_{n+1},\ldots,Z_{2n},Z_1,\ldots,Z_n,Z_{2n+1},Z_{2n+2},\ldots) \in A]$$

$$\vartriangle [(Z_{n+1},\ldots,Z_{2n},Z_1,\ldots,Z_n) \in H \times R^n]\big).$$

But if $A \in \mathscr{S}_{2n}$, this is in turn the same as $P(A \vartriangle V)$. Therefore, $P(A \vartriangle U) = P(A \vartriangle V)$.

But then, $P(A \vartriangle V) < \epsilon$ and $P(A \vartriangle (U \cap V)) \le P(A \vartriangle U) + P(A \vartriangle V) < 2\epsilon$. Since U and V have the same probability and are independent, it follows that $P(A)$ is within ϵ of $P(U)$ and hence $P^2(A)$ is within 2ϵ of $P^2(U) = P(U)P(V) = P(U \cap V)$, which is in turn within 2ϵ of $P(A)$. Therefore, $|P^2(A) - P(A)| < 4\epsilon$ for all ϵ, and so $P(A)$ must be 0 or 1. ■

PROBLEMS

36.1. ↑ Suppose that $[X_t : t \in T]$ is a stochastic process on $(\Omega, \mathscr{F}, P)$ and $A \in \mathscr{F}$. Show that there is a countable subset S of T for which $P[A\|X_t, t \in T] = P[A\|X_t, t \in S]$ with probability 1. Replace A by a random variable and prove a similar result.

36.2. Let T be arbitrary and let $K(s,t)$ be a real function over $T \times T$. Suppose that K is symmetric in the sense that $K(s,t) = K(t,s)$ and nonnegative-definite in the sense that $\sum_{i,j=1}^k K(t_i, t_j) x_i x_j \ge 0$ for $k \ge 1$, $t_1,\ldots,t_k$ in T, and $x_1,\ldots,x_k$ real. Show that there exists a process $[X_t : t \in T]$ for which $(X_{t_1},\ldots,X_{t_k})$ has the centered normal distribution with covariances $K(t_i, t_j)$, $i,j = 1,\ldots,k$.

36.3. Let L be a Borel set on the line, let $\mathscr{L}$ consist of the Borel subsets of L, and let L^T consist of all maps from T into L. Define the appropriate notion of cylinder, and let $\mathscr{L}^T$ be the σ-field generated by the cylinders. State a version of Theorem 36.1 for $(L^T, \mathscr{L}^T)$. Assume T countable, and prove this theorem not by imitating the previous proof but by observing that L^T is a subset of R^T and lies in $\mathscr{R}^T$.

36.4. Suppose that the random variables $X_1, X_2, \ldots$ assume the values 0 and 1 and $P[X_n = 1 \text{ i.o.}] = 1$. Let μ be the distribution over $(0, 1]$ of $\sum_{n=1}^{\infty} X_n / 2^n$. Show that on the unit interval with the measure μ, the digits of the nonterminating dyadic expansion form a stochastic process with the same finite-dimensional distributions as $X_1, X_2, \ldots$.

36.5. 36.3↑ There is an infinite-dimensional version of Fubini's theorem. In the construction in Problem 36.3, let $L = I = (0,1)$, $T = \{1, 2, \ldots\}$, let $\mathscr{I}$ consist of the Borel subsets of I, and suppose that each k-dimensional distribution is the k-fold product of Lebesgue measure over the unit interval. Then I^T is a countable product of copies of $(0,1)$, its elements are sequences $x = (x_1, x_2, \ldots)$ of points of $(0,1)$, and Kolmogorov's theorem ensures the existence on $(I^T, \mathscr{I}^T)$ of a *product* probability measure π: $\pi[x: x_i \leq \alpha_i, i \leq n] = \alpha_1 \cdots \alpha_n$ for $0 \leq \alpha_i \leq 1$. Let I^n denote the n-dimensional unit cube.

(a) Define $\psi: I^n \times I^T \to I^T$ by

$$\psi((x_1, \ldots, x_n),(y_1, y_2, \ldots)) = (x_1, \ldots, x_n, y_1, y_2, \ldots).$$

Show that ψ is measurable $\mathscr{I}^n \times \mathscr{I}^T / \mathscr{I}^T$ and ψ^{-1} is measurable $\mathscr{I}^T / \mathscr{I}^n \times \mathscr{I}^T$. Show that $\psi^{-1}(\lambda_n \times \pi) = \pi$, where λ_n is n-dimensional Lebesgue measure restricted to I^n.

(b) Let f be a function measurable $\mathscr{I}^T$ and, for simplicity, bounded. Define

$$f_n(x_{n+1}, x_{n+2}, \ldots) = \int_0^1 \cdots \int_0^1 f(y_1, \ldots, y_n, x_{n+1}, \ldots) \, dy_1 \cdots dy_n;$$

in other words, integrate out the coordinates one by one. Show by Problem 34.18, martingale theory, and the zero–one law that

$$(36.30) \qquad f_n(x_{n+1}, x_{n+2}, \ldots) \to \int_{I^T} f(y)\pi(dy)$$

except for x in a set of π-measure 0.

(c) Adopting the point of view of part (a), let $g_n(x_1, \ldots, x_n)$ be the result of integrating the variable $(y_{n+1}, y_{n+2}, \ldots)$ out (with respect to π) from $f(x_1, \ldots, x_n, y_{n+1}, \ldots)$. This may suggestively be written as

$$g_n(x_1, \ldots, x_n) = \int_0^1 \int_0^1 \cdots f(x_1, \ldots, x_n, y_{n+1}, y_{n+2}, \ldots) \, dy_{n+1} \, dy_{n+2} \cdots.$$

Show that $g_n(x_1, \ldots, x_n) \to f(x_1, x_2, \ldots)$ except for x in a set of π-measure 0.

36.6. (a) Let T be an interval of the line. Show that $\mathscr{R}^T$ fails to contain the sets of: linear functions, polynomials, constants, nondecreasing functions, functions of bounded variation, differentiable functions, analytic functions, functions con-

tinuous at a fixed t_0, Borel measurable functions. Show that it fails to contain the set of functions that: vanish somewhere in T, satisfy $x(s) < x(t)$ for some pair with $s < t$, have a local maximum anywhere, fail to have a local maximum. (b) Let C be the set of continuous functions on $T = [0, \infty)$. Show that $A \in \mathscr{R}^\Gamma$ and $A \subset C$ imply that $A = \varnothing$. Show, on the other hand, that $A \in \mathscr{R}^T$ and $C \subset A$ do not imply that $A = R^T$.

36.7. Not all systems of finite-dimensional distributions can be realized by stochastic processes for which Ω is the unit interval. Show that there is on the unit interval with Lebesgue measure no process $[X_t: t \geq 0]$ for which the X_t are independent and assume the values 0 and 1 with probability $\frac{1}{2}$ each. Compare Problem 1.1.

36.8. Here is an application of the existence theorem in which T is not a subset of the line. Let $(N, \mathscr{N}, \nu)$ be a measure space, and take T to consist of the $\mathscr{N}$-sets of finite ν-measure. The problem is to construct a generalized Poisson process, a stochastic process $[X_A: A \in T]$ such that (i) X_A has the Poisson distribution with mean $\nu(A)$ and (ii) $X_{A_1}, \ldots, X_{A_n}$ are independent if $A_1, \ldots, A_n$ are disjoint. *Hint:* To define the finite-dimensional distributions, generalize this construction: For A, B in T, consider independent random variables Y_1, Y_2, Y_3 having Poisson distributions with means $\nu(A \cap B^c)$, $\nu(A \cap B)$, $\nu(A^c \cap B)$, take $\mu_{A,B}$ to be the distribution of $(Y_1 + Y_2, Y_2 + Y_3)$.

SECTION 37. BROWNIAN MOTION

Definition

A *Brownian motion* or *Wiener process* is a stochastic process $[W_t: t \geq 0]$, on some $(\Omega, \mathscr{F}, P)$, with these three properties:

(i) *The process starts at* 0:

$$(37.1) \qquad\qquad P[W_0 = 0] = 1.$$

(ii) *The increments are independent*: If

$$(37.2) \qquad\qquad 0 \leq t_0 \leq t_1 \leq \cdots \leq t_k,$$

then

$$(37.3) \qquad P\big[W_{t_i} - W_{t_{i-1}} \in H_i, i \leq k\big] = \prod_{i \leq k} P\big[W_{t_i} - W_{t_{i-1}} \in H_i\big].$$

(iii) *For* $0 \leq s < t$ *the increment* $W_t - W_s$ *is normally distributed with mean* 0 *and variance* $t - s$:

$$(37.4) \qquad P[W_t - W_s \in H] = \frac{1}{\sqrt{2\pi(t-s)}} \int_H e^{-x^2/2(t-s)} \, dx.$$

The existence of such processes will be proved.

Imagine suspended in a fluid a particle bombarded by molecules in thermal motion. The particle will perform a seemingly random movement

first described by the nineteenth-century botanist Robert Brown. Consider a single component of this motion—imagine it projected on a vertical axis—and denote by W_t the height at time t of the particle above a fixed horizontal plane. Condition (i) is merely a convention: the particle starts at 0. Condition (ii) reflects a kind of lack of memory. The displacements $W_{t_1} - W_{t_0}, \ldots, W_{t_{k-1}} - W_{t_{k-2}}$ the particle undergoes during the intervals $[t_0, t_1], \ldots, [t_{k-2}, t_{k-1}]$ in no way influence the displacement $W_{t_k} - W_{t_{k-1}}$ it undergoes during $[t_{k-1}, t_k]$. Although the future behavior of the particle depends on its present position, it does not depend on how the particle got there. As for (iii), that $W_t - W_s$ has mean 0 reflects the fact that the particle is as likely to go up as to go down—there is no drift. The variance grows as the length of the interval $[s, t]$; the particle tends to wander away from its position at time s, and having done so suffers no force tending to restore it to that position. To Norbert Wiener are due the mathematical foundations of the theory of this kind of random motion.
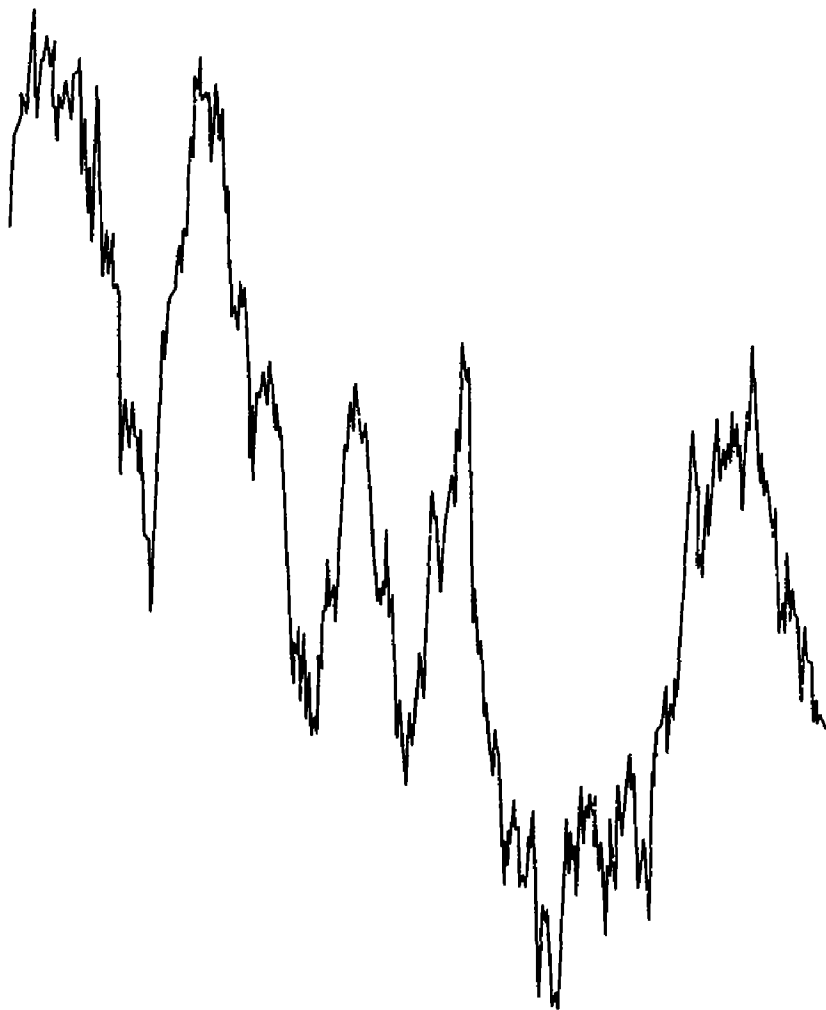
A Brownian motion path.

The increments of the Brownian motion process are *stationary* in the sense that the distribution of $W_t - W_s$ depends only on the difference $t - s$. Since $W_0 = 0$, the distribution of these increments is described by saying that

W_t is normally distributed with mean 0 and variance t. This implies (37.1). If $0 \le s \le t$, then by the independence of the increments, $E[W_s W_t] = E[(W_s(W_t - W_s)] + E[W_s^2] = E[W_s]E[W_t - W_s] + E[W_s^2] = s$. This specifies all the means, variances, and covariances:

$$(37.5) \qquad E[W_t] = 0, \qquad E[W_t^2] = t, \qquad E[W_s W_t] = \min\{s, t\}.$$

If $0 < t_1 < \cdots < t_k$, the joint density of $(W_{t_1}, W_{t_2} - W_{t_1}, \ldots, W_{t_k} - W_{t_{k-1}})$ is by (20.25) the product of the corresponding normal densities. By the Jacobian formula (20.20), $(W_{t_1}, \ldots, W_{t_k})$ has density

$$(37.6) \qquad f_{t_1 \ldots t_k}(x_1, \ldots, x_k) = \prod_{i=1}^{k} \frac{1}{\sqrt{2\pi(t_i - t_{i-1})}} \exp\left[-\frac{(x_i - x_{i-1})^2}{2(t_i - t_{i-1})}\right],$$

where $t_0 = x_0 = 0$.

Sometimes W_t will be denoted $W(t)$, and its value at ω will be $W(t, \omega)$. The nature of the path functions $W(\cdot, \omega)$ will be of great importance.

The existence of the Brownian motion process follows from Kolmogorov's theorem. For $0 < t_1 < \cdots < t_k$ let $\mu_{t_1 \ldots t_k}$ be the distribution in R^k with density (37.6). To put it another way, let $\mu_{t_1 \ldots t_k}$ be the distribution of $(S_1, \ldots, S_k)$, where $S_i = X_1 + \cdots + X_i$ and where $X_1, \ldots, X_k$ are independent, normally distributed random variables with mean 0 and variances $t_1, t_2 - t_1, \ldots, t_k - t_{k-1}$. If $g(x_1, \ldots, x_k) = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_k)$, then $g(S_1, \ldots, S_k) = (S_1, \ldots, S_{i-1}, S_{i+1}, \ldots, S_k)$ has the distribution prescribed for $\mu_{t_1 \ldots t_{i-1} t_{i+1} \ldots t_k}$. This is because $X_i + X_{i+1}$ is normally distributed with mean 0 and variance $t_{i+1} - t_{i-1}$; see Example 20.6. Therefore,

$$(37.7) \qquad \mu_{t_1 \ldots t_{i-1} t_{i+1} \ldots t_k} = \mu_{t_1 \ldots t_k} g^{-1}.$$

The $\mu_{t_1 \ldots t_k}$ defined in this way for increasing, positive $t_1, \ldots, t_k$ thus satisfy the conditions for Kolmogorov's existence theorem as modified in Example 36.4; (37.7) is the same thing as (36.19). Therefore, there does exist a process $[W_t : t > 0]$ corresponding to the $\mu_{t_1 \ldots t_k}$. Taking $W_t = 0$ for $t = 0$ shows that there exists on some $(\Omega, \mathscr{F}, P)$ a process $[W_t : t \ge 0]$ with the finite-dimensional distributions specified by the conditions (i), (ii), and (iii).

Continuity of Paths

If the Brownian motion process is to represent the motion of a particle, it is natural to require that the path functions $W(\cdot, \omega)$ be continuous. But Kolmogorov's theorem does not guarantee continuity. Indeed, for $T = [0, \infty)$, the space $(\Omega, \mathscr{F})$ in the proof of Kolmogorov's theorem is $(R^T, \mathscr{R}^T)$, and as shown in the last section, the set of continuous functions does not lie in $\mathscr{R}^T$.

A special construction gets around this difficulty. The idea is to use for dyadic rational t the random variables W_t as already defined and then to redefine the other W_t in such a way as to ensure continuity. To carry this through requires proving that with probability 1 the sample path is uniformly continuous for dyadic rational arguments in bounded intervals.

Fix a space $(\Omega, \mathscr{F}, P)$ and on it a process $[W_t: t \geq 0]$ having the finite-dimensional distributions prescribed for Brownian motion. Let D be the set of nonnegative dyadic rationals, let $I_{nk} = [k2^{-n}, (k + 2)2^{-n}]$, and put

$$
(37.8) \quad
\begin{aligned}
M_{nk}(\omega) &= \sup_{r \in I_{nk} \cap D} |W(r, \omega) - W(k2^{-n}, \omega)| \\
M_n(\omega) &= \max_{0 \leq k < n2^n} M_{nk}(\omega).
\end{aligned}
$$

Suppose it is shown that $\sum P[M_n > n^{-1}]$ converges. The first Borel–Cantelli lemma will then imply that $B = [M_n > n^{-1} \text{ i.o.}]$ has probability 0. But suppose ω lies outside B. Then for every t and ϵ there exists an n such that $t < n$, $2n^{-1} < \epsilon$, and $M_n(\omega) \leq n^{-1}$. Take $\delta = 2^{-n}$. Suppose that r and r' are dyadic rationals in $[0, t]$ and $|r - r'| < \delta$. Then r and r' must for some $k < n2^n$ lie in a common interval I_{nk} (length 2×2^{-n}), in which case $|W(r, \omega) - W(r', \omega)| \leq 2M_{nk}(\omega) \leq 2M_n(\omega) \leq 2n^{-1} < \epsilon$. Therefore, $\omega \notin B$ implies that $W(r, \omega)$ is for every t uniformly continuous as r ranges over the dyadic rationals in $[0, t]$, and hence it will have a continuous extension to $[0, \infty)$.

To prove $\sum P[M_n > n^{-1}] < \infty$, use Etemadi's maximal inequality (22.10), which applies because of the independence of the increments. This, together with Markov's inequality, gives

$$
\begin{aligned}
&P\left[\max_{i \leq 2^m} |W(t + \delta i 2^{-m}) - W(t)| > \alpha \right] \\
&\leq 3 \max_{i \leq 2^m} P\left[|W(t + \delta i 2^{-m}) - W(t)| \geq \alpha/3 \right] \\
&\leq \frac{3}{(\alpha/3)^4} E\left[(W(t + \delta) - W(t))^4 \right] = \frac{3^5}{\alpha^4} \cdot 3\delta^2 = \frac{K\delta^2}{\alpha^4}
\end{aligned}
$$

(see (21.7) for the moments of the normal distribution). The sets on the left here increase with m, and letting $m \to \infty$ leads to

$$
(37.9) \quad P\left[\sup_{\substack{0 \leq r \leq 1 \\ r \in D}} |W(t + r\delta) - W(t)| > \alpha \right] \leq \frac{K\delta^2}{\alpha^4}.
$$

Therefore,

$$
P[M_n > n^{-1}] \leq n2^n \frac{K(2 \times 2^{-n})^2}{(n^{-1})^4} = \frac{4Kn^5}{2^n},
$$

and $\sum P[M_n > n^{-1}]$ does converge.

Therefore, there exists a measurable set B such that $P(B) = 0$ and such that for ω outside B, $W(r, \omega)$ is uniformly continuous as r ranges over the dyadic rationals in any bounded interval. If $\omega \notin B$ and r decreases to t through dyadic rational values, then $W(r, \omega)$ has the Cauchy property and hence converges. Put

$$W_t'(\omega) = W'(t, \omega) = \begin{cases} \lim_{r \downarrow t} W(r, \omega) & \text{if } \omega \notin B, \\ 0 & \text{if } \omega \in B, \end{cases}$$

where r decreases to t through the set D of dyadic rationals. By construction, $W'(t, \omega)$ is continuous in t for each ω in Ω. If $\omega \notin B$, then $W(r, \omega) = W'(r, \omega)$ for dyadic rationals, and $W'(\cdot, \omega)$ is the continuous extension to all of $[0, \infty)$.

The next thing is to show that the W_t' have the same joint distributions as the W_t. It is convenient to prove this by a lemma which will be used again further on.

Lemma 1.   *Let X_n and X be k-dimensional random vectors, and let $F_n(x)$ be the distribution function of X_n. If $X_n \to X$ with probability 1 and $F_n(x) \to F(x)$ for all x, then $F(x)$ is the distribution function of X.*

PROOF.[†]   Let X have distribution function H. By two applications of Theorem 4.1, if $h > 0$, then

$$F(x_1, \ldots, x_k) = \limsup_n F_n(x_1, \ldots, x_k) \le H(x_1, \ldots, x_k)$$

$$\le \liminf_n F_n(x_1 + h, \ldots, x_k + h)$$

$$= F(x_1 + h, \ldots, x_k + h).$$

It follows by continuity from above that F and H agree.   ∎

Now, for $0 < t_1 < \cdots < t_k$, choose dyadic rationals $r_i(n)$ decreasing to the t_i. Apply Lemma 1 with $(W_{r_1(n)}, \ldots, W_{r_k(n)})$ and $(W_{t_1}', \ldots, W_{t_k}')$ in the roles of X_n and X, and with the distribution function with density (37.6) in the role of F. Since (37.6) is continuous in the t_i, it follows by Scheffé's theorem that $F_n(x) \to F(x)$, and by construction $X_n \to X$ with probability 1. By the lemma, $(W_{t_1}', \ldots, W_{t_k}')$ has distribution function F, which of course is also the distribution function of $(W_{t_1}, \ldots, W_{t_k})$.

Thus $[W_t' : t \ge 0]$ is a stochastic process, on the same probability space as $[W_t, t \ge 0]$, which has the finite-dimensional distributions required for Brownian motion and moreover has a continuous sample path $W'(\cdot, \omega)$ for every ω.

By enlarging the set B in the definition of $W'_t(\omega)$ to include all the ω for which $W(0, \omega) \neq 0$, one can also ensure that $W'(0, \omega) = 0$. Now discard the original random variables W_t and relabel W'_t as W_t. The new $[W_t : t \geq 0]$ is a stochastic process satisfying conditions (i), (ii), and (iii) for Brownian motion and this one as well:

(iv) *For each ω, $W(t, \omega)$ is continuous in t and $W(0, \omega) = 0$.*

From now on, by a Brownian motion will be meant a process satisfying (iv) as well as (i), (ii), and (iii). What has been proved is this:

Theorem 37.1. *There exist processes $[W_t : t \geq 0]$ satisfying conditions* (i), (ii), (iii), *and* (iv)—*Brownian motion processes.*

In the construction above, W_r for dyadic r was used to define W_t in general. For that reason it suffices to apply Kolmogorov's theorem for a countable index set. By the second proof of that theorem, the space $(\Omega, \mathcal{F}, P)$ can be taken as the unit interval with Lebesgue measure.

The next section treats a general scheme for dealing with path-function questions by in effect replacing an uncountable time set by a countable one.

Measurable Processes

Let T be a Borel set on the line, let $[X_t : t \in T]$ be a stochastic process on an $(\Omega, \mathcal{F}, P)$, and consider the mapping

$$(37.10) \qquad\qquad (t, \omega) \to X_t(\omega) = X(t, \omega)$$

carrying $T \times \Omega$ into R^1. Let $\mathcal{T}$ be the σ-field of Borel subsets of T. The process is said to be *measurable* if the mapping (37.10) is measurable $\mathcal{T} \times \mathcal{F} / \mathcal{R}^1$.

In the presence of measurability, each sample path $X(\cdot, \omega)$ is measurable $\mathcal{T}$ by Theorem 18.1. Then, for example, $\int_a^b \varphi(X(t, \omega)) \, dt$ makes sense if $(a, b) \subset T$ and φ is a Borel function, and by Fubini's theorem

$$E\left[\int_a^b \varphi(X(t, \cdot)) \, dt \right] = \int_a^b E[\varphi(X_t)] \, dt \quad \text{if} \quad \int_a^b E[|\varphi(X_t)|] \, dt < \infty.$$

Hence the usefulness of this result:

Theorem 37.2. *Brownian motion is measurable.*

PROOF. If

$$W^{(n)}(t, \omega) = W(k 2^{-n}, \omega) \qquad \text{for} \qquad k 2^{-n} \leq t < (k + 1) 2^{-n},$$
$$k = 0, 1, 2, \ldots,$$

then the mapping $(t, \omega) \to W^{(n)}(t, \omega)$ is measurable $\mathcal{T} \times \mathcal{F}$. But by the continuity of the sample paths, this mapping converges to the mapping (37.10) pointwise (for every (t, ω)), and so by Theorem 13.4(ii) the latter mapping is also measurable $\mathcal{T} \times \mathcal{F}/\mathcal{R}^1$. ∎

Irregularity of Brownian Motion Paths

Starting with a Brownian motion $[W_t: t \geq 0]$ define

$$(37.11) \qquad W_t'(\omega) = c^{-1} W_{c^2 t}(\omega),$$

where $c > 0$. Since $t \to c^2 t$ is an increasing function, it is easy to see that the process $[W_t': t \geq 0]$ has independent increments. Moreover, $W_t' - W_s' = c^{-1}(W_{c^2 t} - W_{c^2 s})$, and for $s \leq t$ this is normally distributed with mean 0 and variance $c^{-2}(c^2 t - c^2 s) = t - s$. Since the paths $W'(\cdot, \omega)$ all start from 0 and are continuous, $[W_t': t \geq 0]$ is another Brownian motion. In (37.11) the time scale is contracted by the factor c^2, but the other scale only by the factor c.

That the transformation (37.11) preserves the properties of Brownian motion implies that the paths, although continuous, must be highly irregular. It seems intuitively clear that for c large enough the path $W(\cdot, \omega)$ must with probability nearly 1 have somewhere in the time interval $[0, c]$ a chord with slope exceeding, say, 1. But then $W'(\cdot, \omega)$ has in $[0, c^{-1}]$ a chord with slope exceeding c. Since the W_t' are distributed as the W_t, this makes it plausible that $W(\cdot, \omega)$ must in arbitrarily small intervals $[0, \delta]$ have chords with arbitrarily great slopes, which in turn makes it plausible that $W(\cdot, \omega)$ cannot be differentiable at 0. More generally, mild irregularities in the path will become ever more extreme under the transformation (37.11) with ever larger values of c. It is shown below that, in fact, the paths are with probability 1 nowhere differentiable.

Also interesting in this connection is the transformation

$$(37.12) \qquad W_t''(\omega) = \begin{cases} t W_{1/t}(\omega) & \text{if } t > 0, \\ 0 & \text{if } t = 0. \end{cases}$$

Again it is easily checked that the increments are independent and normally distributed with the means and variances appropriate to Brownian motion. Moreover, the path $W''(\cdot, \omega)$ is continuous except possibly at $t = 0$. But (37.9) holds with W_s'' in place of W_s because it depends only on the finite-dimensional distributions, and by the continuity of $W''(\cdot, \omega)$ over $(0, \infty)$ the supremum is the same if not restricted to dyadic rationals. Therefore, $P[\sup_{s \leq n^{-3}} |W_s''| > n^{-1}] \leq K/n^2$, and it follows by the first Borel–Cantelli lemma that $W''(\cdot, \omega)$ is continuous also at 0 for ω outside a set M of probability 0. For $\omega \in M$, redefine $W''(t, \omega) \equiv 0$; then $[W_t'': t \geq 0]$ is a Brownian motion and (37.12) holds with probability 1.

The behavior of $W(\cdot, \omega)$ near 0 can be studied through the behavior of $W''(\cdot, \omega)$ near ∞ and vice versa. Since $(W_t'' - W_0'')/t = W_{1/t}$, $W''(\cdot, \omega)$ cannot have a derivative at 0 if $W(\cdot, \omega)$ has no limit at ∞. Now, in fact,

$$(37.13) \qquad \inf_n W_n = -\infty, \qquad \sup_n W_n = +\infty$$

with probability 1. To prove this, note that $W_n = X_1 + \cdots + X_n$, where the $X_k = W_k - W_{k-1}$ are independent. Consider

$$\left[\sup_n W_n < \infty\right] = \bigcup_{u=1}^{\infty} \bigcap_{m=1}^{\infty} \left[\max_{i \le m} W_i \le u\right];$$

this is a tail set and hence by the zero–one law has probability 0 or 1. Now $-X_1, -X_2, \ldots$ have the same joint distributions as $X_1, X_2, \ldots$, and so this event has the same probability as

$$\left[\inf_n W_n > -\infty\right] = \bigcup_{u=1}^{\infty} \bigcap_{m=1}^{\infty} \left[\max_{i \le m} (-W_i) \le u\right].$$

If these two sets have probability 1, so has $[\sup_n |W_n| < \infty]$, so that $P[\sup_n |W_n| < x] > 0$ for some x. But $P[|W_n| < x] = P[|W_1| < x/n^{1/2}] \to 0$. This proves (37.13).

Since (37.13) holds with probability 1, $W''(\cdot, \omega)$ has with probability 1 upper and lower right derivatives of $+\infty$ and $-\infty$ at $t = 0$. The same must be true of every Brownian motion. A similar argument shows that, for each fixed t, $W(\cdot, \omega)$ is nondifferentiable at t with probability 1. In fact, $W(\cdot, \omega)$ is nowhere differentiable:

Theorem 37.3. *For ω outside a set of probability 0, $W(\cdot, \omega)$ is nowhere differentiable.*

PROOF. The proof is direct—makes no use of the transformations (37.11) and (37.12). Let

$$(37.14) \qquad X_{nk} = \max\left\{\left|W\left(\frac{k+1}{2^n}\right) - W\left(\frac{k}{2^n}\right)\right|, \left|W\left(\frac{k+2}{2^n}\right) - W\left(\frac{k+1}{2^n}\right)\right|, \right.$$
$$\left.\left|W\left(\frac{k+3}{2^n}\right) - W\left(\frac{k+2}{2^n}\right)\right|\right\}.$$

By independence and the fact that the differences here have the distribution of $2^{-n/2} W_1$, $P[X_{nk} \le \epsilon] = P^3[|W_1| \le 2^{n/2}\epsilon]$; since the standard normal density is bounded by 1, $P[X_{nk} \le \epsilon] \le (2 \times 2^{n/2}\epsilon)^3$. If $Y_n = \min_{k \le n2^n} X_{nk}$, then

$$(37.15) \qquad P[Y_n \le \epsilon] \le n2^n (2 \times 2^{n/2}\epsilon)^3.$$

Consider now the upper and lower right-hand derivatives

$$D^W(t, \omega) = \limsup_{h \downarrow 0} \frac{W(t+h, \omega) - W(t, \omega)}{h},$$

$$D_W(t, \omega) = \liminf_{h \downarrow 0} \frac{W(t+h, \omega) - W(t, \omega)}{h}.$$

Define E (not necessarily in $\mathscr{F}$) as the set of ω such that $D^W(t, \omega)$ and $D_W(t, \omega)$ are both finite for some value of t. Suppose that ω lies in E, and suppose specifically that

$$-K < D_W(t, \omega) \leq D^W(t, \omega) < K.$$

There exists a positive δ (depending on ω, t, and K) such that $t \leq s \leq t + \delta$ implies $|W(s, \omega) - W(t, \omega)| \leq K|s - t|$. If n exceeds some n_0 (depending on δ, K, and t), then

$$4 \times 2^{-n} < \delta, \qquad 8K < n, \qquad n > t.$$

Given such an n, choose k so that $(k-1)2^{-n} \leq t < k2^{-n}$. Then $|i2^{-n} - t| < \delta$ for $i = k, k+1, k+2, k+3$, and therefore $X_{nk}(\omega) \leq 2K(4 \times 2^{-n}) < n2^{-n}$. Since $k - 1 \leq t2^n < n2^n$, $Y_n(\omega) \leq n2^{-n}$.

What has been shown is that if ω lies in E, then ω lies in $A_n = [Y_n \leq n2^{-n}]$ for all sufficiently large n: $E \subset \liminf_n A_n$. By (37.15),

$$P(A_n) \leq n2^n(2 \times 2^{n/2} \times n2^{-n})^3 \to 0.$$

By Theorem 4.1, $\liminf_n A_n$ has probability 0, and outside this set $W(\cdot, \omega)$ is nowhere differentiable—in fact, nowhere does it have finite upper and lower right-hand derivatives. (Similarly, outside a set of probability 0, nowhere does $W(\cdot, \omega)$ have finite upper and lower left-hand derivatives.) ∎

If A is the set of ω for which $W(\cdot, \omega)$ has a derivative somewhere, what has been shown is that $A \subset B$ for a measurable B such that $P(B) = 0$; $P(A) = 0$ if A is measurable, but this has not been proved. To avoid such problems in the study of continuous-time processes, it is convenient to work in a *complete* probability space. The space $(\Omega, \mathscr{F}, P)$ is complete (see p. 44) if $A \subset B$, $B \in \mathscr{F}$, and $P(B) = 0$ together imply that $A \in \mathscr{F}$ (and then, of course, $P(A) = 0$). If the space is not already complete, it is possible to enlarge $\mathscr{F}$ to a new σ-field and extend P to it in such a way that the new space is complete. The following assumption therefore entails no loss of generality: *For the rest of this section the space $(\Omega, \mathscr{F}, P)$ on which the Brownian motion is defined is assumed complete.* Theorem 37.3 now becomes: $W(\cdot, \omega)$ is with probability 1 nowhere differentiable.

A nowhere-differentiable path represents the motion of a particle that at no time has a velocity. Since a function of bounded variation is differentiable almost everywhere (Section 31), $W(\cdot, \omega)$ is of unbounded variation with probability 1. Such a path represents the motion of a particle that in its wanderings back and forth travels an infinite distance in finite time. The Brownian motion model thus does not in its fine structure represent physical reality. The irregularity of the Brownian motion paths is of considerable mathematical interest, however. A continuous, nowhere-differentiable function is regarded as pathological, or used to be, but from the Brownian-motion point of view such functions are the rule not the exception.[†]

The set of zeros of the Brownian motion is also interesting. By property (iv), $t = 0$ is a zero of $W(\cdot, \omega)$ for each ω. Now $[W_t''': t \ge 0]$ as defined by (37.12) is another Brownian motion, and so by (37.13) the sequence $\{W_n''': n = 1, 2, \ldots\} = \{nW_{1/n}: n = 1, 2, \ldots\}$ has supremum $+\infty$ and infimum $-\infty$ for ω outside a set of probability 0; for such an ω, $W(\cdot, \omega)$ changes sign infinitely often near 0 and hence by continuity has zeros arbitrarily near 0. Let $Z(\omega)$ denote the set of zeros of $W(\cdot, \omega)$. What has just been shown is that $0 \in Z(\omega)$ for each ω and that 0 is with probability 1 a limit of positive points in $Z(\omega)$. From (37.13) it also follows that $Z(\omega)$ is with probability 1 unbounded above. More is true:

Theorem 37.4. *The set $Z(\omega)$ is with probability 1 perfect [A15], unbounded, nowhere dense, and of Lebesgue measure 0.*

PROOF. Since $W(\cdot, \omega)$ is continuous, $Z(\omega)$ is closed for every ω. Let λ denote Lebesgue measure. Since Brownian motion is measurable (Theorem 37.2), Fubini's theorem applies:

$$\int_\Omega \lambda(Z(\omega)) P(d\omega) = (\lambda \times P)[(t, \omega): W(t, \omega) = 0]$$

$$= \int_0^\infty P[\omega: W(t, \omega) = 0] \, dt = 0.$$

Thus $\lambda(Z(\omega)) = 0$ with probability 1.

If $W(\cdot, \omega)$ is nowhere differentiable, it cannot vanish throughout an interval I and hence must by continuity be nonzero throughout some subinterval of I. By Theorem 37.3, then, $Z(\omega)$ is with probability 1 nowhere dense.

It remains to show that each point of $Z(\omega)$ is a limit of other points of $Z(\omega)$. As observed above, this is true of the point 0 of $Z(\omega)$. For the general point of $Z(\omega)$, a stopping-time argument is required. Fix $r \ge 0$ and let

---

[†]For the construction of a specific example, see Problem 31.18.

$\tau(\omega) = \inf[t: t \geq r, W(t, \omega) = 0]$; note that this set is nonempty with probability 1 by (37.13). Thus $\tau(\omega)$ is the first zero following r. Now

$$\left[\omega: \tau(\omega) \leq t\right] = \left[\omega: \inf_{r \leq s \leq t} |W(s, \omega)| = 0\right],$$

and by continuity the infimum here is unchanged if s is restricted to rationals. This shows that τ is a random variable and that

$$\left[\omega. \tau(\omega) \leq t\right] \in \sigma[W_u: u \leq t].$$

A nonnegative random variable with this property is a *stopping time*.

To know the value of τ is to know at most the values of W_u for $u \leq \tau$. Since the inciements are independent, it therefore seems intuitively clear that the process

$$(37.16) \quad W_t^*(\omega) = W_{\tau(\omega)+t}(\omega) - W_{\tau(\omega)}(\omega) = W_{\tau(\omega)+t}(\omega), \qquad t \geq 0,$$

ought itself to be a Brownian motion. This is, in fact, true by the next result, Theorem 37.5. What is proved there is that the finite-dimensional distributions of $[W_t^*: t \geq 0]$ are the right ones for Brownian motion. The other properties are obvious: $W^*(\cdot, \omega)$ is continuous and vanishes at 0 by construction, and the space on which $[W_t^*; t \geq 0]$ is defined is complete because it is the original space $(\Omega, \mathscr{F}, P)$, assumed complete.

If $[W_t^*: t \geq 0]$ is indeed a Brownian motion, then, as observed above, for ω outside a set B_r of probability 0 there is a positive sequence $\{t_n\}$ such that $t_n \to 0$ and $W^*(t_n, \omega) = 0$. But then $W(\tau(\omega) + t_n, \omega) = 0$, so that $\tau(\omega)$, a zero of $W(\cdot, \omega)$, is the limit of other larger zeros of $W(\cdot, \omega)$. Now $\tau(\omega)$ was the first zero following r. (There is a different stopping time τ for each r, but the notation does not show this.) If B is the union of the B_r for rational r, the first point of $Z(\omega)$ following r is a limit of other, larger points of $Z(\omega)$. Suppose that $\omega \notin B$ and $t \in Z(\omega)$, where $t > 0$; it is to be shown that t is a limit of other points of $Z(\omega)$. If t is the limit of smaller points of $Z(\omega)$, there is of course nothing to prove. Otherwise, there is a rational r such that $r < t$ and $W(\cdot, \omega)$ does not vanish in $[r, t)$; but then, since $\omega \notin B_r$, t is a limit of larger points s that lie in $Z(\omega)$. This completes the proof of Theorem 37.4 under the provisional assumption that (37.16) is a Brownian motion.   ∎

The Strong Markov Property

Fix $t_0 \geq 0$ and put

$$(37.17) \qquad\qquad W_t' = W_{t_0+t} - W_{t_0}, \qquad t \geq 0.$$

It is easily checked that $[W_t': t \geq 0]$ has the finite-dimensional distributions

appropriate to Brownian motion. As the other properties are obvious, it is in fact a Brownian motion.

Let

(37.18) $$\mathscr{F}_t = \sigma[W_s : s \le t].$$

The random variables (37.17) are independent of $\mathscr{F}_{t_0}$. To see this, suppose that $0 \le s_1 \le \cdots \le s_j \le t_0$ and $0 \le t_1 \le \cdots \le t_k$. Put $u_i = t_0 + t_i$. Since the increments are independent, $(W'_{t_1}, W'_{t_2} - W'_{t_1}, \ldots, W'_{t_k} - W'_{t_{k-1}}) = (W_{u_1} - W_{t_0}, W_{u_2} - W_{u_1}, \ldots, W_{u_k} - W_{u_{k-1}})$ is independent of $(W_{s_1}, W_{s_2} - W_{s_1}, \ldots, W_{s_j} - W_{s_{j-1}})$. But then $(W'_{t_1}, W'_{t_2}, \ldots, W'_{t_k})$ is independent of $(W_{s_1}, W_{s_2}, \ldots, W_{s_j})$. By Theorem 4.2, $(W'_{t_1}, \ldots, W'_{t_k})$ is independent of $\mathscr{F}_{t_0}$. Thus

(37.19) $$P\big(\big[(W'_{t_1}, \ldots, W'_{t_k}) \in H\big] \cap A\big)$$

$$= P\big[(W'_{t_1}, \ldots, W'_{t_k}) \in H\big] P(A)$$

$$= P\big[(W_{t_1}, \ldots, W_{t_k}) \in H\big] P(A), \qquad A \in \mathscr{F}_{t_0},$$

where the second equality follows because (37.17) is a Brownian motion. This holds for all H in $\mathscr{R}^k$.

The problem now is to prove all this when t_0 is replaced by a *stopping time* τ—a nonnegative random variable for which

(37.20) $$[\omega : \tau(\omega) \le t] \in \mathscr{F}_t, \qquad t \ge 0.$$

It will be assumed that τ is finite, at least with probability 1. Since $[\tau = t] = [\tau \le t] - \bigcup_n [\tau \le t - n^{-1}]$, (37.20) implies that

(37.21) $$[\omega : \tau(\omega) = t] \in \mathscr{F}_t, \qquad t \ge 0.$$

The conditions (37.20) and (37.21) are analogous to the conditions (7.18) and (35.18), which prevent prevision on the part of the gambler.

Now $\mathscr{F}_{t_0}$ contains the information on the past of the Brownian motion up to time t_0, and the analogue for τ is needed. Let $\mathscr{F}_\tau$ consist of all measurable sets M for which

(37.22) $$M \cap [\omega : \tau(\omega) \le t] \in \mathscr{F}_t$$

for all t. (See (35.20) for the analogue in discrete time.) Note that $\mathscr{F}_\tau$ is a σ-field and τ is measurable $\mathscr{F}_\tau$. Since $M \cap [\tau = t] = M \cap [\tau = t] \cap [\tau \le t]$,

(37.23) $$M \cap [\omega : \tau(\omega) = t] \in \mathscr{F}_t$$

for M in $\mathscr{F}_\tau$. For example, $\tau = \inf[t : W_t = 1]$ is a stopping time and $[\inf_{s \le \tau} W_s > -1]$ is in $\mathscr{F}_\tau$.

Theorem 37.5. *Let τ be a stopping time, and put*

$$(37.24) \qquad W_t^*(\omega) = W_{\tau(\omega)+t}(\omega) - W_{\tau(\omega)}(\omega), \qquad t \geq 0.$$

Then $[W_t^*: t \geq]$ is a Brownian motion, and it is independent of $\mathscr{F}_\tau$—that is, $\sigma[W_t^*: t \geq 0]$ is independent of $\mathscr{F}_\tau$:

$$(37.25) \quad P\Big(\big[(W_{t_1}^*, \ldots, W_{t_k}^*) \in H\big] \cap M\Big)$$

$$= P\big[(W_{t_1}^*, \ldots, W_{t_k}^*) \in H\big] P(M) = P\big[(W_{t_1}, \ldots, W_{t_k}) \in H\big] P(M)$$

for H in $\mathscr{R}^k$ and M in $\mathscr{F}_\tau$.

That the transformation (37.24) preserves Brownian motion is the *strong Markov property*.[†] Part of the conclusion is that the W_t^* are random variables.

PROOF. Suppose first that τ has countable range V and let t_0 be the general point of V. Since

$$\big[\omega: W_t^*(\omega) \in H\big] = \bigcup_{t_0 \in V} \big[\omega: W_{t_0+t}(\omega) - W_{t_0}(\omega) \in H, \tau(\omega) = t_0\big],$$

W_t^* is a random variable. Also,

$$P\Big(\big[(W_{t_1}^*, \ldots, W_{t_k}^*) \in H\big] \cap M\Big)$$

$$= \sum_{t_0 \in V} P\Big(\big[(W_{t_1}^*, \ldots, W_{t_k}^*) \in H\big] \cap M \cap [\tau = t_0]\Big).$$

If $M \in \mathscr{F}_\tau$, then $M \cap [\tau = t_0] \in \mathscr{F}_{t_0}$ by (37.23). Further, if $\tau = t_0$, then W_t^* coincides with W_t' as defined by (37.17). Therefore, (37.19) reduces this last sum to

$$\sum_{t_0 \in V} P\big[(W_{t_1}, \ldots, W_{t_k}) \in H\big] P(M \cap [\tau = t_0])$$

$$= P\big[(W_{t_1}, \ldots, W_{t_k}) \in H\big] P(M).$$

This proves the first and third terms in (37.25) equal; to prove equality with the middle term, simply consider the case $M = \Omega$.

[†]Since the Brownian motion has independent increments, it is a Markov process (see Examples 33.9 and 33.10); hence the terminology.

Thus the theorem holds if τ has countable range. For the general τ, put

$$(37.26) \qquad \tau_n = \begin{cases} k2^{-n} & \text{if } (k-1)2^{-n} < \tau \le k2^{-n}, \ k = 1,2,\ldots \\ 0 & \text{if } \tau = 0. \end{cases}$$

If $k2^{-n} \le t < (k+1)2^{-n}$, then $[\tau_n \le t] = [\tau \le k2^{-n}] \in \mathscr{F}_{k2^{-n}} \subset \mathscr{F}_t$. Thus each τ_n is a stopping time. Suppose that $M \in \mathscr{F}_\tau$ and $k2^{-n} \le t < (k+1)2^{-n}$. Then $M \cap [\tau_n \le t] = M \cap [\tau \le k2^{-n}] \in \mathscr{F}_{k2^{-n}} \subset \mathscr{F}_t$. Thus $\mathscr{F}_\tau \subset \mathscr{F}_{\tau_n}$. Let $W_t^{(n)}(\omega) = W_{\tau_n(\omega)+t}(\omega) - W_{\tau_n(\omega)}(\omega)$ —that is, let $W_t^{(n)}$ be the W_t^* corresponding to the stopping time τ_n. If $M \in \mathscr{F}_\tau$ then $M \in \mathscr{F}_{\tau_n}$, and by an application of (37.25) to the discrete case already treated,

$$(37.27) \qquad P\left(\left[\left(W_{t_1}^{(n)},\ldots,W_{t_k}^{(n)}\right) \in H\right] \cap M\right) = P\left[\left(W_{t_1},\ldots,W_{t_k}\right) \in H\right] P(M).$$

But $\tau_n(\omega) \to \tau(\omega)$ for each ω, and by continuity of the sample paths, $W_t^{(n)}(\omega) \to W_t^*(\omega)$ for each ω. Condition on M and apply Lemma 1 with $(W_{t_1}^{(n)},\ldots,W_{t_k}^{(n)})$ for X_n, $(W_{t_1}^*,\ldots,W_{t_k}^*)$ for X, and the distribution function of $(W_{t_1},\ldots,W_{t_n})$ for $F = F_n$. Then (37.25) follows from (37.27). ∎

The τ in the proof of Theorem 37.4 is a stopping time, and so (37.16) is a Brownian motion, as required in that proof. Further applications will be given below.

If $\mathscr{F}^* = \sigma[W_t^*: t \ge 0]$, then according to (37.25) (and Theorem 4.2) the σ-fields $\mathscr{F}_\tau$ and $\mathscr{F}^*$ are independent:

$$(37.28) \qquad P(A \cap B) = P(A)P(B), \qquad A \in \mathscr{F}_\tau, \quad B \in \mathscr{F}^*.$$

For fixed t define τ_n by (37.26) but with $t2^{-n}$ in place of 2^{-n} at each occurrence. Then $[W_\tau < x] \cap [\tau \le t]$ is the limit superior of the sets $[W_{\tau_n} < x] \cap [\tau \le t]$, each of which lies in $\mathscr{F}_t$. This proves that $[W_\tau < x]$ lies in $\mathscr{F}_\tau$ and hence that W_τ is measurable $\mathscr{F}_\tau$. Since τ is measurable $\mathscr{F}_\tau$,

$$(37.29) \qquad [(\tau, W_\tau) \in H] \in \mathscr{F}_\tau$$

for planar Borel sets H.

The Reflection Principle

For a stopping time τ, define

$$(37.30) \qquad W_t'' = \begin{cases} W_t & \text{if } t \le \tau, \\ W_\tau - (W_t - W_\tau) & \text{if } t \ge \tau. \end{cases}$$

The sample path for $[W_t'': t \ge 0]$ is the same as the sample path for $[W_t: t \ge 0]$ up to τ, and beyond that it is reflected through the point W_τ. See the figure.
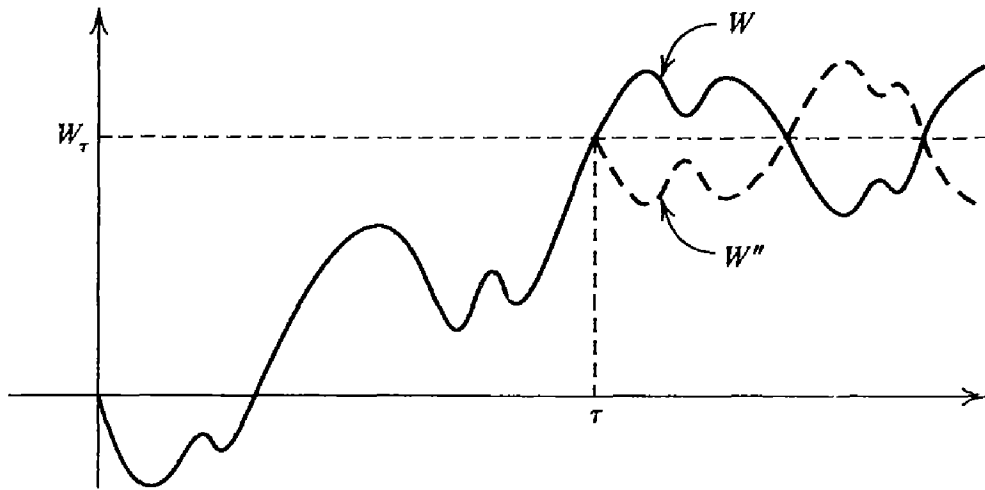
The process defined by (37.30) is a Brownian motion, and to prove it, one need only check the finite-dimensional distributions: $P[(W_{t_1}, \ldots, W_k) \in H] = P[(W''_{t_1}, \ldots, W''_{t_k}) \in H]$. By the argument starting with (37.26), it is enough to consider the case where τ has countable range, and for this it is enough to check the equation when the sets are intersected with $[\tau = t_0]$.

Consider for notational simplicity a pair of points:

$$(37.31) \qquad P\big[\tau = t_0, (W_s, W_t) \in H\big] = P\big[\tau = t_0, (W''_s, W''_t) \in H\big].$$

If $s \le t \le t_0$, this holds because the two events are identical. Suppose next that $s \le t_0 \le t$. Since $[\tau = t_0]$ lies in $\mathscr{F}_{t_0}$, it follows by the independence of the increments, symmetry, and the definition (37.30) that

$$P\big[\tau = t_0, \big(W_s, W_{t_0}\big) \in I, W_t - W_{t_0} \in J\big]$$

$$= P\big[\tau = t_0, \big(W_s, W_{t_0}\big) \in I, -\big(W_t - W_{t_0}\big) \in J\big]$$

$$= P\big[\tau = t_0, \big(W''_s, W''_{t_0}\big) \in I, W''_t - W''_{t_0} \in J\big].$$

If $K = I \times J$, this is

$$P\big[\tau = t_0, \big(W_s, W_{t_0}, W_t - W_{t_0}\big) \in K\big] = P\big[\tau = t_0, \big(W''_s, W''_{t_0}, W''_t - W''_{t_0}\big) \in K\big],$$

and by π–λ it follows for all $K \in \mathscr{R}^3$. For the appropriate K, this gives (37.31). The remaining case, $t_0 \le s \le t$, is similar.

These ideas can be used to derive in a very simple way the distribution of $M_t = \sup_{s \le t} W_s$. Suppose that $x > 0$. Let $\tau = \inf[s: W_s \ge x]$, define W'' by (37.30), and put $\tau'' = \inf[s: W''_s \ge x]$ and $M''_t = \sup_{s \le t} W''_s$. Since $\tau'' = \tau$ and W'' is another Brownian motion, reflection through the point $W_\tau = x$ shows

that

$$P[M_t \geq x] = P[\tau \leq t]$$

$$= P[\tau \leq t, W_t \leq x] + P[\tau \leq t, W_t \geq x]$$

$$= P[\tau'' \leq t, W_t'' \leq x] + P[\tau \leq t, W_t \geq x]$$

$$= P[\tau'' \leq t, W_t \geq x] + P[\tau \leq t, W_t \geq x]$$

$$= P[\tau \leq t, W_t \geq x] + P[\tau \leq t, W_t \geq x] = 2P[W_t \geq x].$$

Therefore,

$$(37.32) \qquad P[M_t \geq x] = \frac{2}{\sqrt{2\pi}} \int_{x/\sqrt{t}}^{\infty} e^{-u^2/2} \, du.$$

This argument, an application of the *reflection principle*,[†] becomes quite transparent when referred to the diagram.

Skorohod Embedding*

Suppose that $X_1, X_2, \ldots$ are independent and identically distributed random variables with mean 0 and variance σ^2. A powerful method, due to Skorohod, of studying the partial sums $S_n = X_1 + \cdots + X_n$ is to construct an increasing sequence $\tau_0 = 0, \tau_1, \tau_2, \ldots$ of stopping times such that $W(\tau_n)$ has the same distribution as S_n. The differences $\tau_k - \tau_{k-1}$ will turn out to be independent and identically distributed with mean σ^2, so that by the law of large numbers $n^{-1}\tau_n = n^{-1}\sum_{k=1}^{n}(\tau_k - \tau_{k-1})$ is likely to be near σ^2. But if τ_n is near $n\sigma^2$, then by the continuity of Brownian motion paths $W(\tau_n)$ will be near $W(n\sigma^2)$, and so the distribution of $S_n/\sigma\sqrt{n}$, which coincides with the distribution of $W(\tau_n)/\sigma\sqrt{n}$, will be near the distribution of $W(n\sigma^2)/\sigma\sqrt{n}$ —that is, will be near the standard normal distribution. The method will thus yield another proof of the central limit theorem, one independent of the characteristic-function arguments of Section 27.

But it will also give more. For example, the distribution of $\max_{k \leq n} S_k/\sigma\sqrt{n}$ is exactly the distribution of $\max_{k \leq n} W(\tau_k)/\sigma\sqrt{n}$, and this in turn is near the distribution of $\sup_{t \leq n\sigma^2} W(t)/\sigma\sqrt{n}$, which can be written down explicitly because of (37.32). It will thus be possible to derive the limiting distribution of $\max_{k \leq n} S_k$. The joint behavior of the partial sums is closely related to the behavior of Brownian motion paths.

The Skorohod construction involves the class $\mathcal{T}$ of stopping times for which

$$(37.33) \qquad\qquad E[W_\tau] = 0,$$

$$(37.34) \qquad\qquad E[\tau] = E[W_\tau^2],$$

[†]See Problem 37.18 for another application.
*The rest of this section, which requires martingale theory, may be omitted.

and

(37.35) $$E[\tau^2] \le 4E[W_\tau^4].$$

Lemma 2. *All bounded stopping times are members of $\mathcal{T}$.*

PROOF. Define $Y_{\theta,t} = \exp(\theta W_t - \frac{1}{2}\theta^2 t)$ for all θ and for $t \ge 0$. Suppose that $s \le t$ and $A \in \mathcal{F}_s$. Since Brownian motion has independent increments,

$$\int_A Y_{\theta,t}\, dP = \int_A e^{\theta W_s - \theta^2 s/2}\, dP \cdot E\left[e^{\theta(W_t - W_s) - \theta^2(t-s)/2}\right],$$

and a calculation with moment generating functions (see Example 21.2) shows that

(37.36) $$\int_A Y_{\theta,s}\, dP = \int_A Y_{\theta,t}\, dP, \qquad s \le t, \quad A \in \mathcal{F}_s.$$

This says that for θ fixed, $[Y_{\theta,t}: t \ge 0]$ is a continuous-time martingale adapted to the σ-fields $\mathcal{F}_t$. It is the *moment-generating-function martingale* associated with the Brownian motion.

Let $f(\theta, t)$ denote the right side of (37.36). By Theorem 16.8,

$$\frac{\partial}{\partial\theta} f(\theta, t) = \int_A Y_{\theta,t}(W_t - \theta t)\, dP,$$

$$\frac{\partial^2}{\partial\theta^2} f(\theta, t) = \int_A Y_{\theta,t}\left[(W_t - \theta t)^2 - t\right] dP,$$

$$\frac{\partial^4}{\partial\theta^4} f(\theta, t) = \int_A Y_{\theta,t}\left[(W_t - \theta t)^4 - 6(W_t - \theta t)^2 t + 3t^2\right] dP.$$

Differentiate the other side of the equation (37.36) the same way and set $\theta = 0$. The result is

$$\int_A W_s\, dP = \int_A W_t\, dP, \qquad\qquad s \le t, \quad A \in \mathcal{F}_s,$$

$$\int_A (W_s^2 - s)\, dP = \int_A (W_t^2 - t)\, dP, \qquad\qquad s \le t, \quad A \in \mathcal{F}_s,$$

$$\int_A (W_s^4 - 6W_s^2 s + 3s^2)\, dP = \int_A (W_t^4 - 6W_t^2 t + 3t^2)\, dP, \qquad s \le t, \quad A \in \mathcal{F}_s.$$

This gives three more martingales: If Z_t is any of the three random variables

(37.37) $$W_t, \qquad W_t^2 - t, \qquad W_t^4 - 6W_t^2 t + 3t^2,$$

then $Z_0 = 0$, Z_t is integrable and measurable $\mathscr{F}_t$, and

(37.38) $$\int_A Z_s \, dP = \int_A Z_t \, dP, \qquad s \le t, \quad A \in \mathscr{F}_s.$$

In particular, $E[Z_t] = [Z_0] = 0$.

If τ is a stopping time with finite range $\{t_1, \ldots, t_m\}$ bounded by t, then (37.38) implies that

$$E[Z_\tau] = \sum_i \int_{[\tau = t_i]} Z_{t_i} \, dP = \sum_i \int_{[\tau = t_i]} Z_t \, dP = E[Z_t] = 0.$$

Suppose that τ is bounded by t but does not necessarily have finite range. Put $\tau_n = k 2^{-n} t$ if $(k-1) 2^{-n} t < \tau \le k 2^{-n} t$, $1 \le k \le 2^n$, and put $\tau_n = 0$ if $\tau = 0$. Then τ_n is a stopping time and $E[Z_{\tau_n}] = 0$. For each of the three possibilities (37.37) for Z_t, $\sup_{s \le t} |Z_s|$ is integrable because of (37.32). It therefore follows by the dominated convergence theorem that $E[Z_\tau] = \lim_n E[Z_{\tau_n}] = 0$.

Thus $E[Z_\tau] = 0$ for every bounded stopping time τ. The three cases (37.37) give

$$E[W_\tau] = E[W_\tau^2 - \tau] = E[W_\tau^4 - 6W_\tau^2 \tau + 3\tau^2] = 0.$$

This implies (37.33), (37.34), and

$$0 = E[W_\tau^4] - 6E[W_\tau^2 \tau] + 3E[\tau^2]$$
$$\ge E[W_\tau^4] - 6E^{1/2}[W_\tau^4] E^{1/2}[\tau^2] + 3E[\tau^2].$$

If $C = E^{1/2}[W_\tau^4]$ and $x = E^{1/2}[\tau^2]$, the inequality is $0 \ge q(x) = 3x^2 - 6Cx + C^2$. Each zero of q is at most $2C$, and q is negative only between these two zeros. Therefore, $x \le 2C$, which implies (37.35).   ∎

Lemma 3. *Suppose that τ and τ_n are stopping times, that each τ_n is a member of $\mathscr{F}$, and that $\tau_n \to \tau$ with probability 1. Then τ is a member of $\mathscr{F}$ if* (i) $E[W_{\tau_n}^4] \le E[W_\tau^4] < \infty$ *for all n, or if* (ii) *the $W_{\tau_n}^4$ are uniformly integrable.*

PROOF. Since Brownian motion paths are continuous, $W_{\tau_n} \to W_\tau$ with probability 1. Each of the two hypotheses (i) and (ii) implies that $E[W_{\tau_n}^4]$ is bounded and hence that $E[\tau_n^2]$ is bounded, and it follows (see (16.28)) that the sequences $\{\tau_n\}$, $\{W_{\tau_n}\}$, and $\{W_{\tau_n}^2\}$ are uniformly integrable. Hence (37.33) and (37.34) for τ follow by Theorem 16.14 from the same relations for the τ_n. The first hypothesis implies that $\liminf_n E[W_{\tau_n}^4] \le E[W_\tau^4]$, and the second implies that $\lim_n E[W_{\tau_n}^4] = E[W_\tau^4]$. In either case it follows by Fatou's lemma that $E[\tau^2] \le \liminf_n E[\tau_n^2] \le 4 \liminf_n E[W_{\tau_n}^4] \le 4E[W_\tau^4]$.   ∎

Suppose that $a, b \geq 0$ and $a + b > 0$, and let $\tau(a, b)$ be the *hitting time* for the set $\{-a, b\}$: $\tau(a, b) = \inf[t: W_t \in \{-a, b\}]$. By (37.13), $\tau(a, b)$ is finite with probability 1, and it is a stopping time because $\tau(a, b) \leq t$ if and only if for every m there is a rational $r \leq t$ for which W_r is within m^{-1} of $-a$ or of b. From $|W(\min\{\tau(a, b), n\})| \leq \max\{a, b\}$ it follows by Lemma 3(ii) that $\tau(a, b)$ is a member of $\mathcal{T}$. Since $W_{\tau(a, b)}$ assumes only the values $-a$ and b, $E[W_{\tau(a, b)}] = 0$ implies that

$$(37.39) \quad P\big[W_{\tau(a, b)} = -a\big] = \frac{b}{a + b}, \qquad P\big[W_{\tau(a, b)} = b\big] = \frac{a}{a + b}.$$

This is obvious on grounds of symmetry in the case $a = b$.

Let μ be a probability measure on the line with mean 0. The program is to construct a stopping time τ for which W_τ has distribution μ. Assume that $\mu\{0\} < 1$, since otherwise $\tau \equiv 0$ obviously works. If μ consists of two point masses, they must for some positive a and b be a mass of $b/(a + b)$ at $-a$ and a mass of $a/(a + b)$ at b; in this case $\tau_{(a, b)}$ is by (37.39) the required stopping time. The general case will be treated by adding together stopping times of this sort.

Consider a random variable X having distribution μ. (The probability space for X has nothing to do with the space the given Brownian motion is defined on.) The technique will be to represent X as the limit of a martingale $X_1, X_2, \ldots$ of a simple form and then to duplicate the martingale by $W_{\tau_1}, W_{\tau_2}, \ldots$ for stopping times τ_n; the τ_n will have a limit τ such that W_τ has the same distribution as X.

The first step is to construct sets

$$\Delta_n: a_0^{(n)} < a_1^{(n)} < \cdots < a_{r_n}^{(n)}$$

and corresponding partitions

$$\mathcal{P}_n: \begin{cases} I_0^n = \left(-\infty, a_0^{(n)}\right], \\ I_k^n = \left(a_{k-1}^{(n)}, a_k^{(n)}\right], \ 1 \leq k \leq r_n, \\ I_{r_n + 1}^n = \left(a_{r_n}^{(n)}, \infty\right). \end{cases}$$

Let $M(H)$ be the conditional mean:

$$M(H) = \frac{1}{\mu(H)} \int_H x \mu(dx) \qquad \text{if } \mu(H) > 0.$$

Let Δ_1 consist of the single point $M(R^1) = E[X] = 0$, so that $\mathcal{P}_1$ consists of $I_0^1 = (-\infty, 0]$ and $I_1^1 = (0, \infty)$. Suppose that Δ_n and $\mathcal{P}_n$ are given. If $\mu((I_k^n)^\circ) > 0$, split I_k^n by adding to Δ_n the point $M(I_k^n)$, which lies in $(I_k^n)^\circ$; if $\mu((I_k^n)^\circ) = 0$, I_k^n appears again in $\mathcal{P}_{n+1}$.

Let $\mathscr{S}_n$ be the σ-field generated by the sets $[X \in I_k^n]$, and put $X_n = E[X \| \mathscr{S}_n]$. Then $X_1, X_2, \ldots$ is a martingale and $X_n = M(I_k^n)$ on $[X \in I_k^n]$. The X_n have finite range, and their joint distributions can be written out explicitly. In fact, $[X_1 = M(I_{k_1}^1), \ldots, X_n = M(I_{k_n}^n)] = [X \in I_{k_1}^1, \ldots, X \in I_{k_n}^n]$, and this set is empty unless $I_{k_1}^1 \supset \cdots \supset I_{k_n}^n$, in which case it is $[X_n = M(I_{k_n}^n)] = [X \in I_{k_n}^n]$. Therefore, if $k_{n-1} = j$ and $I_j^{n-1} = I_{k-1}^n \cup I_k^n$,

$$P\left[X_n = M(I_{k-1}^n) \| X_1 = M(I_{k_1}^1), \ldots, X_{n-1} = M(I_{k_{n-1}}^{n-1})\right] = \frac{\mu(I_{k-1}^n)}{\mu(I_j^{n-1})}$$

and

$$P\left[X_n = M(I_k^n) \| X_1 = M(I_{k_1}^1), \ldots, X_{n-1} = M(I_{k_{n-1}}^{n-1})\right] = \frac{\mu(I_k^n)}{\mu(I_j^{n-1})},$$

provided the conditioning event has positive probability. Thus the martingale $\{X_n\}$ has the Markov property, and if $x = M(I_j^{n-1})$, $u = M(I_{k-1}^n)$, and $v = M(I_k^n)$, then the conditional distribution of X_n given $X_{n-1} = x$ is concentrated at the two points u and v and has mean x. The structure of $\{X_n\}$ is determined by these conditional probabilities together with the distribution

$$P\left[X_1 = M(I_0^1)\right] = \mu(I_0^1), \qquad P\left[X_1 = M(I_1^1)\right] = \mu(I_1^1).$$

of X_1.

If $\mathscr{S} = \sigma(\bigcup_n \mathscr{S}_n)$, then $X_n \to E[X \| \mathscr{S}]$ with probability 1 by the martingale theorem (Theorem 35.6). But, in fact, $X_n \to X$ with probability 1, as the following argument shows. Let B be the union of all open sets of μ-measure 0. Then B is a countable disjoint union of open intervals; enlarge B by adding to it any endpoints of μ-measure 0 these intervals may have. Then $\mu(B) = 0$, and $x \notin B$ implies that $\mu(x - \epsilon, x] > 0$ and $\mu[x, x + \epsilon) > 0$ for all positive ϵ. Suppose that $x = X(\omega) \notin B$ and let $x_n = X_n(\omega)$. Let $I_{k_n}^n$ be the element of $\mathscr{S}_n$ containing x; then $x_{n+1} = M(I_{k_n}^n)$ and $I_{k_n}^n \downarrow I$ for some interval I. Suppose that $x_{n+1} < x - \epsilon$ for n in an infinite sequence N of integers. Then x_{n+1} is the left endpoint of $I_{k_{n+1}}^{n+1}$ for n in N and converges along N to the left endpoint, say a, of I, and $(x - \epsilon, x] \subset I$. Further, $x_{n+1} = M(I_{k_n}^n) \to M(I)$ along N, so that $M(I) = a$. But this is impossible because $\mu(x - \epsilon, x] > 0$. Therefore, $x_n \geq x - \epsilon$ for large n. Similarly, $x_n \leq x + \epsilon$ for large n, and so $x_n \to x$. Thus $X_n(\omega) \to X(\omega)$ if $X(\omega) \notin B$, the probability of which is 1.

Now $X_1 = E[X \| \mathscr{S}_1]$ has mean 0, and its distribution consists of point masses at $-a = M(I_0^1)$ and $b = M(I_1^1)$. If $\tau_1 = \tau(a, b)$ is the hitting time to

$\{-a, b\}$, then (see (37.39)) τ_1 is a stopping time, a member of $\mathscr{T}$, and W_{τ_1} has the same distribution as X_1.

Let τ_2 be the infimum of those t for which $t \geq \tau_1$ and W_t is one of the points $M(I_k^2)$, $0 \leq k \leq r_2 + 1$. By (37.13), τ_2 is finite with probability 1; it is a stopping time, because $\tau_2 \leq t$ if and only if for every m there are rationals r and s such that $r \leq s + m^{-1}$, $r \leq t$, $s \leq t$, W_r is within m^{-1} of one of the points $M(I_j^1)$, and W_s is within m^{-1} of one of the points $M(I_k^2)$. Since $|W(\min\{\tau_2, n\})|$ is at most the maximum of the values $|M(I_k^2)|$, it follows by Lemma 3(ii) that τ_2 is a member of $\mathscr{T}$.

Define W_t^* by (37.24) with τ_1 for τ. If $x = M(I_j^1)$, then x is an endpoint common to two adjacent intervals I_{k-1}^2 and I_k^2; put $u = M(I_{k-1}^2)$ and $v = M(I_k^2)$. If $W_{\tau_1} = x$, then u and v are the only possible values of W_{τ_2}. If τ^* is the first time the Brownian motion $[W_t^*: t \geq 0]$ hits $u - x$ or $v - x$, then by (37.39),

$$P[W_{\tau^*}^* = u - x] = \frac{v - x}{v - u}, \qquad P[W_{\tau^*}^* = v - x] = \frac{x - u}{v - u}.$$

On the set $[W_{\tau_1} = x]$, τ_2 coincides with $\tau_1 + \tau^*$, and it follows by (37.28) that

$$P\big[W_{\tau_1} = x, W_{\tau_2} = v\big] = P\big[W_{\tau_1} = x, x + W_{\tau^*}^* = v\big]$$

$$= P\big[W_{\tau_1} = x\big] P[W_{\tau^*}^* = v - x] = P\big[W_{\tau_1}\big] \frac{x - u}{v - u}.$$

This, together with the same computation with u in place of v, shows that for $W_{\tau_1} = x$ the conditional distribution of W_{τ_2} is concentrated at the two points u and v and has mean x. Thus the conditional distribution of W_{τ_2} given W_{τ_1} coincides with the conditional distribution of X_2 given X_1. Since W_{τ_1} and X_1 have the same distribution, the random vectors (W_{τ_1}, W_{τ_2}) and (X_1, X_2) also have the same distribution.

An inductive extension of this argument proves the existence of a sequence of stopping times τ_n such that $\tau_1 \leq \tau_2 \leq \cdots$, each τ_n is a member of $\mathscr{T}$, and for each n, $W_{\tau_1}, \ldots, W_{\tau_n}$ have the same joint distribution as $X_1, \ldots, X_n$. Now suppose that X has finite variance. Since τ_n is a member of $\mathscr{T}$, $E[\tau_n] = E[W_{\tau_n}^2] = E[X_n^2] = E[E^2[X \| \mathscr{G}_n]] \leq E[X^2]$ by Jensen's inequality (34.7). Thus $\tau = \lim_n \tau_n$ is finite with probability 1. Obviously it is a stopping time, and by path continuity, $W_{\tau_n} \to W_\tau$ with probability 1. Since $X_n \to X$ with probability 1, it is a consequence of the following lemma that W_τ has the distribution of X.

Lemma 4. *If $X_n \to X$ and $Y_n \to Y$ with probability 1, and if X_n and Y_n have the same distribution, then so do X and Y.*

PROOF.[†]   By two applications of (4.9),

$$P[X \le x] \le P[X < x + \epsilon] \le \lim_n \inf P[X_n \le x + \epsilon]$$

$$\le \lim_n \sup P[Y_n \le x + \epsilon] \le P[Y \le x + \epsilon].$$

Let $\epsilon \to 0$: $P[X \le x] \le P[Y \le x]$. Now interchange the roles of X and Y.   ∎

Since $X_n^2 \le E[X^2 \| \mathscr{G}_n]$, the X_n are uniformly integrable by the lemma preceding Theorem 35.6. By the monotone convergence theorem and Theorem 16.14, $E[\tau] = \lim_n E[\tau_n] = \lim_n E[W_{\tau_n}^2] = \lim_n E[X_n^2] = E[X^2] = E[W_\tau^2]$. If $E[X^4] < \infty$, then $E[W_{\tau_n}^4] = E[X_n^4] \le E[X^4] = E[W_\tau^4]$ (Jensen's inequality again), and so τ is a member of $\mathscr{T}$. Hence $E[\tau^2] \le 4E[W_\tau^4]$.

This construction establishes the first of Skorohod's embedding theorems:

Theorem 37.6.   *Suppose that X is a random variable with mean 0 and finite variance. There is a stopping time τ such that W_τ has the same distribution as X, $E[\tau] = E[X^2]$, and $E[\tau^2] \le 4E[X^4]$.*

Of course, the last inequality is trivial unless $E[X^4]$ is finite. The theorem could be stated in terms not of X but of its distribution, the point being that the probability space X is defined on is irrelevant. Skorohod's second embedding theorem is this:

Theorem 37.7.   *Suppose that $X_1, X_2, \ldots$ are independent and identically distributed random variables with mean 0 and finite variance, and put $S_n = X_1 + \cdots + X_n$. There is a nondecreasing sequence $\tau_1, \tau_2, \ldots$ of stopping times such that the W_{τ_n} have the same joint distributions as the S_n and $\tau_1, \tau_2 - \tau_1, \tau_3 - \tau_2, \ldots$ are independent and identically distributed random variables satisfying $E[\tau_n - \tau_{n-1}] = E[X_1^2]$ and $E[(\tau_n - \tau_{n-1})^2] \le 4E[X_1^4]$.*

PROOF.   The method is to repeat the construction above inductively. For notational clarity write $W_t = W_t^{(1)}$ and put $\mathscr{F}_t^{(1)} = \sigma[W_s^{(1)}: 0 \le s \le t]$ and $\mathscr{F}^{(1)} = \sigma[W_t^{(1)}: t \ge 0]$. Let δ_1 be the stopping time of Theorem 37.6, so that $W_{\delta_1}^{(1)}$ and X_1 have the same distribution. Let $\mathscr{F}_{\delta_1}^{(1)}$ be the class of M such that $M \cap [\delta_1 \le t] \in \mathscr{F}_t^{(1)}$ for all t.

Now put $W_t^{(2)} = W_{\delta_1 + t}^{(1)} - W_{\delta_1}^{(1)}$, $\mathscr{F}_t^{(2)} = \sigma[W_s^{(2)}: 0 \le s \le t]$, and $\mathscr{F}^{(2)} = \sigma[W_t^{(2)}: t \ge 0]$. By another application of Theorem 37.6, construct a stopping time δ_2 for the Brownian motion $[W_t^{(2)}: t \ge 0]$ in such a way that $W_{\delta_2}^{(2)}$ has the same distribution as X_1. In fact, use for δ_2 the very same martingale construction as for δ_1, so that $(\delta_1, W_{\delta_1}^{(1)})$ and $(\delta_2, W_{\delta_2}^{(2)})$ have the same distribution. Since $\mathscr{F}_{\delta_1}^{(1)}$ and $\mathscr{F}^{(2)}$ are independent (see (37.28)), it follows (see (37.29)) that $(\delta_1, W_{\delta_1}^{(1)})$ and $(\delta_2, W_{\delta_2}^{(2)})$ are independent.

[†]This is obvious from the weak-convergence point of view.

Let $\mathscr{F}_{\delta_2}^{(2)}$ be the class of M such that $M \cap [\delta_2 \le t] \in \mathscr{F}_t^{(2)}$ for all t. If $W_t^{(3)} = W_{\delta_2 + t}^{(2)} - W_{\delta_2}^{(2)}$ and $\mathscr{F}^{(3)}$ is the σ-field generated by these random variables, then again $\mathscr{F}_{\delta_2}^{(2)}$ and $\mathscr{F}^{(3)}$ are independent. These two σ-fields are contained in $\mathscr{F}^{(2)}$, which is independent of $\mathscr{F}_{\delta_1}^{(1)}$. Therefore, the three σ-fields $\mathscr{F}_{\delta_1}^{(1)}, \mathscr{F}_{\delta_2}^{(2)}, \mathscr{F}^{(3)}$ are independent. The procedure therefore extends inductively to give independent, identically distributed random vectors $(\delta_n, W_{\delta_n}^{(n)})$. If $\tau_n = \delta_1 + \cdots + \delta_n$, then $W_{\tau_n}^{(1)} = W_{\delta_1}^{(1)} + \cdots + W_{\delta_n}^{(n)}$ has the distribution of $X_1 + \cdots + X_n$. ∎

Invariance*

If $E[X_1^2] = \sigma^2$, then, since the random variables $\tau_n - \tau_{n-1}$ of Theorem 37.7 are independent and identically distributed, the strong law of large numbers (Theorem 22.1) applies and hence so does the weak one:

$$(37.40) \qquad P\big[|n^{-1}\tau_n - \sigma^2| \ge \epsilon\big] \to 0.$$

(If $E[X_1^4] < \infty$, so that the $\tau_n - \tau_{n-1}$ have second moments, this follows immediately by Chebyshev's inequality.) Now S_n has the distribution of $W(\tau_n)$, and τ_n is near $n\sigma^2$ by (37.40); hence S_n should have nearly the distribution of $W(n\sigma^2)$, namely the normal distribution with mean 0 and variance $n\sigma^2$.

To prove this, choose an increasing sequence of integers N_k such that $P[|n^{-1}\tau_n - \sigma^2| \ge k^{-1}] < k^{-1}$ for $n \ge N_k$, and put $\epsilon_n = k^{-1}$ for $N_k \le n < N_{k+1}$. Then $\epsilon_n \to 0$ and $P[|n^{-1}\tau_n - \sigma^2| \ge \epsilon_n] < \epsilon_n$. By two applications of (37.32),

$$\delta_n(\epsilon) = P\left[\frac{|W(n\sigma^2) - W(\tau_n)|}{\sigma\sqrt{n}} \ge \epsilon\right]$$

$$\le P\big[|n^{-1}\tau_n - \sigma^2| \ge \epsilon_n\big] + P\left[\sup_{|t - n\sigma^2| \le \epsilon_n n} |W(t) - W(n\sigma^2)| \ge \epsilon\sigma\sqrt{n}\right]$$

$$\le \epsilon_n + 4P\big[|W(\epsilon_n n)| \ge \epsilon\sigma\sqrt{n}\big],$$

and it follows by Chebyshev's inequality that $\lim_n \delta_n(\epsilon) = 0$. Since S_n is distributed as $W(\tau_n)$,

$$P\left[\frac{W(n\sigma^2)}{\sigma\sqrt{n}} \le x - \epsilon\right] - \delta_n(\epsilon) \le P\left[\frac{S_n}{\sigma\sqrt{n}} \le x\right]$$

$$\le P\left[\frac{W(n\sigma^2)}{\sigma\sqrt{n}} \le x + \epsilon\right] + \delta_n(\epsilon).$$

*This topic may be omitted.

Here $W(n\sigma^2)/\sigma\sqrt{n}$ can be replaced by a random variable N with the standard normal distribution, and letting $n \to \infty$ and then $\epsilon \to 0$ shows that

$$\lim_n P\left[\frac{S_n}{\sigma\sqrt{n}} \le x\right] = P[N \le x].$$

This gives a new proof of the central limit theorem for independent, identically distributed random variables with second moments (the Lindeberg–Lévy theorem—Theorem 27.1). Observe that none of the convergence theory of Chapter 5 has been used.

This proof of the central limit theorem is an application of the *invariance principle*: S_n has nearly the distribution of $W(n\sigma^2)$, and the distribution of the latter does not depend on (vary with) the distribution common to the X_n. More can be said if the X_n have fourth moments.

For each n, define a stochastic process $[Y_n(t): 0 \le t \le 1]$ by $Y_n(0, \omega) = 0$ and

$$(37.41) \quad Y_n(t, \omega) = \frac{1}{\sigma\sqrt{n}} S_k(\omega) \quad \text{if } \frac{k-1}{n} < t \le \frac{k}{n}, \qquad k = 1, \dots, n.$$

If $k/n = t > 0$ and n is large, then k is large, too, and $Y_n(t) = t^{1/2}S_k/\sigma\sqrt{k}$ is by the central limit theorem approximately normally distributed with mean 0 and variance t. Since the X_n are independent, the increments of (37.41) should be approximately independent, and so the process should behave approximately as a Brownian motion does.

Let τ_n be the stopping times of Theorem 37.7, and in analogy with (37.41) put $Z_n(0) = 0$ and

$$(37.42) \quad Z_n(t) = \frac{1}{\sigma\sqrt{n}} W(\tau_k) \quad \text{if } \frac{k-1}{n} < t \le \frac{k}{n}, \qquad k = 1, \dots, n.$$

By construction, the finite-dimensional distributions of $[Y_n(t): 0 \le t \le 1]$ coincide with those of $[Z_n(t): 0 \le t \le 1]$. It will be shown that the latter process nearly coincides with $[W(tn\sigma^2)/\sigma\sqrt{n}: 0 \le t \le 1]$, which is itself a Brownian motion over the time interval $[0, 1]$—see (37.11). Put $W_n(t) = W(tn\sigma^2)/\sigma\sqrt{n}$.

Let $B_n(\delta)$ be the event that $|\tau_k - k\sigma^2| \ge \delta n\sigma^2$ for some $k \le n$. By Kolmogorov's inequality (22.9),

$$(37.43) \qquad P(B_n(\delta)) \le \frac{\text{Var}[\tau_n]}{\delta^2 n^2 \sigma^4} \le \frac{4E[X_1^4]}{\delta^2 n \sigma^4} \to 0.$$

If $(k-1)n^{-1} < t \le kn^{-1}$ and $n > \delta^{-1}$, then

$$\left|\frac{\tau_k}{n\sigma^2} - t\right| \le \left|\frac{\tau_k}{n\sigma^2} - \frac{k}{n}\right| + \frac{1}{n} \le 2\delta$$

on the event $(B_n(\delta))^c$, and so

$$\left| Z_n(t) - W_n(t) \right| = \left| W_n\left(\frac{\tau_k}{n\sigma^2}\right) - W_n(t) \right| \le \sup_{|s-t| \le 2\delta} \left| W_n(s) - W_n(t) \right|$$

on $(B_n(\delta))^c$. Since the distribution of this last random variable is unchanged if the $W_n(t)$ are replaced by $W(t)$,

$$P\left[\sup_{t \le 1} |Z_n(t) - W_n(t)| \ge \epsilon \right]$$

$$\le P(B_n(\delta)) + P\left[\sup_{t \le 1} \sup_{|s-t| \le 2\delta} |W(s) - W(t)| \ge \epsilon \right].$$

Let $n \to \infty$ and then $\delta \to 0$; it follows by (37.43) and the continuity of Brownian motion paths that

$$(37.44) \qquad \lim_n P\left[\sup_{t \le 1} |Z_n(t) - W_n(t)| \ge \epsilon \right] = 0$$

for positive ϵ. Since the processes (37.41) and (37.42) have the same finite-dimensional distributions, this proves the following general invariance principle or *functional central limit theorem*.

Theorem 37.8. *Suppose that* $X_1, X_2, \ldots$ *are independent, identically distributed random variables with mean* 0, *variance* σ^2, *and finite fourth moments, and define* $Y_n(t)$ *by* (37.41). *There exist* (*on another probability space*), *for each* n, *processes* $[Z_n(t): 0 \le t \le 1]$ *and* $[W_n(t): 0 \le t \le 1]$ *such that the first has the same finite-dimensional distributions as* $[Y_n(t): 0 \le t \le 1]$, *the second is a Brownian motion, and* $P[\sup_{t \le 1} |Z_n(t) - W_n(t)| \ge \epsilon] \to 0$ *for positive* ϵ.

As an application, consider the maximum $M_n = \max_{k \le n} S_k$. Now $M_n/\sigma\sqrt{n} = \sup_t Y_n(t)$ has the same distribution as $\sup_t Z_n(t)$, and it follows by (37.44) that

$$P\left[\left| \sup_{t \le 1} Z_n(t) - \sup_{t \le 1} W_n(t) \right| \ge \epsilon \right] \to 0.$$

But $P[\sup_{t \le 1} W_n(t) \ge x] = P[\sup_{t \le 1} W(t) \ge x] = 2P[N \ge x]$ for $x \ge 0$ by (37.32). Therefore,

$$(37.45) \qquad P\left[\frac{M_n}{\sigma\sqrt{n}} \le x \right] \to 2P[N \le x], \qquad x \ge 0.$$

PROBLEMS

37.1. 36.2 ↑ Show that $K(s, t) = \min\{s, t\}$ is nonnegative-definite; use Problem 36.2 to prove the existence of a process with the finite-dimensional distributions prescribed for Brownian motion.

37.2. Let $X(t)$ be independent, standard normal variables, one for each dyadic rational t (Theorem 20.4; the unit interval can be used as the probability space). Let $W(0) = 0$ and $W(n) = \sum_{k=1}^{n} X(k)$. Suppose that $W(t)$ is already defined for dyadic rationals of rank n, and put

$$W\left(\frac{2k+1}{2^{n+1}}\right) = \tfrac{1}{2}W\left(\frac{k}{2^n}\right) + \tfrac{1}{2}W\left(\frac{k+1}{2^n}\right) + \frac{1}{2^{1+n/2}}X\left(\frac{2k+1}{2^{n+1}}\right).$$

Prove by induction that the $W(t)$ for dyadic t have the finite-dimensional distributions prescribed for Brownian motion. Now construct a Brownian motion with continuous paths by the argument leading to Theorem 37.1. This avoids an appeal to Kolmogorov's existence theorem.

37.3. ↑ For each n define new variables $W_n(t)$ by setting $W_n(k/2^n) = W(k/2^n)$ for dyadics of order n and interpolating linearly in between. Set $\delta_n = \sup_{t \le n}|W_{n+1}(t) - W_n(t)|$, and show that

$$\delta_n = \max_{0 \le k < n2^n} \left| W\left(\frac{2k+1}{2^{n+1}}\right) - \left[\tfrac{1}{2}W\left(\frac{k}{2^n}\right) + \tfrac{1}{2}W\left(\frac{k+1}{2^n}\right)\right]\right|.$$

The construction in the preceding problem makes it clear that the difference here is normal with variance $1/2^{n+2}$. Find positive x_n such that $\sum x_n$ and $\sum P[\delta_n \ge x_n]$ both converge, and conclude that outside a set of probability 0, $W_n(t, \omega)$ converges uniformly over bounded intervals. Replace $W(t, \omega)$ by $\lim_n W_n(t, \omega)$. This gives another construction of a Brownian motion with continuous paths.

37.4. 36.6 ↑ Let $T = [0, \infty)$, and let P be a probability measure on $(R^T, \mathcal{R}^T)$ having the finite-dimensional distributions prescribed for Brownian motion. Let C consist of the continuous elements of R^T.

(a) Show that $P_*(C) = 0$, or $P^*(R^T - C) = 1$ (see (3.9) and (3.10)). Thus completing $(R^T, \mathcal{R}^T, P)$ will not give C probability 1.

(b) Show that $P^*(C) = 1$.

37.5. Suppose that $[W_t: t \ge 0]$ is some stochastic process having independent, stationary increments satisfying $E[W_t] = 0$ and $E[W_t^2] = t$. Show that if the finite-dimensional distributions are preserved by the transformation (37.11), then they must be those of Brownian motion.

37.6. Show that $\bigcap_{t > 0} \sigma[W_s: s \ge t]$ contains only sets of probability 0 and 1. Do the same for $\bigcap_{\epsilon > 0} \sigma[W_t: 0 < t < \epsilon]$; give examples of sets in this σ-field.

37.7. Show by a direct argument that $W(\cdot, \omega)$ is with probability 1 of unbounded variation on $[0, 1]$: Let $Y_n = \sum_{i=1}^{2^n} |W(i2^{-n}) - W((i-1)2^{-n})|$. Show that Y_n has mean $2^{n/2}E[|W_1|]$ and variance at most $\text{Var}[|W_1|]$. Conclude that $\sum P[Y_n < n] < \infty$.

37.8. Show that the Poisson process as defined by (23.5) is measurable.

37.9. Show that for $T = [0, \infty)$ the coordinate-variable process $[Z_t: t \in T]$ on $(R^T, \mathscr{R}^T)$ is not measurable.

37.10. Extend Theorem 37.4 to the set $[t: W(t, \omega) = \alpha]$.

37.11. Let τ_α be the first time the Brownian motion hits $\alpha > 0$: $\tau_\alpha = \inf[t: W_t \geq \alpha]$. Show that the distribution of τ_α has over $(0, \infty)$ the density

$$(37.46) \qquad h_\alpha(t) = \frac{\alpha}{\sqrt{2\pi}} \frac{1}{t^{3/2}} e^{-\alpha^2/2t}.$$

Show that $E[\tau_\alpha] = \infty$. Show that τ_α has the same distribution as α^2/N^2, where N is a standard normal variable.

37.12. ↑ **(a)** Show by the strong Markov property that τ_α and $\tau_{\alpha+\beta} - \tau_\alpha$ are independent and that the latter has the same distribution as τ_β. Conclude that $h_\alpha * h_\beta = h_{\alpha+\beta}$. Show that $\beta\tau_\alpha$ has the same distribution as $\tau_{\alpha\sqrt{\beta}}$.

(b) Show that each h_α is stable—see Problem 28.10.

37.13. ↑ Suppose that $X_1, X_2, \ldots$ are independent and each has the distribution (37.46).

(a) Show that $(X_1 + \cdots + X_n)/n^2$ also has the distribution (37.46). Contrast this with the law of large numbers.

(b) Show that $P[n^{-2} \max_{k \leq n} X_k \leq x] \to \exp(-\alpha\sqrt{2/\pi x})$ for $x > 0$. Relate this to Theorem 14.3.

37.14. 37.11↑ Let $\rho(s, t)$ be the probability that a Brownian path has at least one zero in (s, t). From (37.46) and the Markov property deduce

$$(37.47) \qquad \rho(s, t) = \frac{2}{\pi} \arccos \sqrt{\frac{s}{t}}.$$

Hint: Condition with respect to W_s.

37.15. ↑ **(a)** Show that the probability of no zero in $(t, 1)$ is $(2/\pi) \arcsin \sqrt{t}$ and hence that the position of the last zero preceding 1 is distributed over $(0, 1)$ with density $\pi^{-1}(t(1 - t))^{-1/2}$.

(b) Similarly calculate the distribution of the position of the first zero following time 1.

(c) Calculate the joint distribution of the two zeros in (a) and (b).

37.16. ↑ **(a)** Show by Theorem 37.8 that $\inf_{s \leq u \leq t} Y_n(u)$ and $\inf_{s \leq u \leq t} Z_n(u)$ both converge in distribution to $\inf_{s \leq u \leq t} W(u)$ for $0 \leq s \leq t \leq 1$. Prove a similar result for the supremum.

(b) Let $A_n(s, t)$ be the event that S_k, the position at time k in a symmetric random walk, is 0 for at least one k in the range $sn \leq k \leq tn$, and show that $P(A_n(s, t)) \to (2/\pi) \arccos \sqrt{s/t}$.

(c) Let T_n be the maximum k such that $k \le n$ and $S_k = 0$. Show that T_n/n has asymptotically the distribution with density $\pi^{-1}(t(1-t))^{-1/2}$ over $(0,1)$. As this density is larger at the ends of the interval than in the middle, the last time during a night's play a gambler was even is more likely to be either early or late than to be around midnight.

37.17. ↑ Show that $\rho(s,t) = \rho(t^{-1}, s^{-1}) = \rho(cs, ct)$. Check this by (37.47) and also by the fact that the transformations (37.11) and (37.12) preserve the properties of Brownian motion.

37.18. Deduce by the reflection principle that (M_t, W_t) has density

$$\frac{2(2y-x)}{t\sqrt{2\pi t}} \exp\left[-\frac{(2y-x)^2}{2t} \right]$$

on the set where $y \ge 0$ and $y \ge x$. Now deduce from Theorem 37.8 the corresponding limit theorem for symmetric random walk.

37.19. Show by means of the transformation (37.12) that for positive a and b the probability is 1 that the process is within the boundary $-at < W_t < bt$ for all sufficiently large t. Show that $a/(a+b)$ is the probability that it last touches above rather than below.

37.20. The martingale calculation used for (37.39) also works for slanting boundaries. For positive a, b, r, let τ be the smallest t such that either $W_t = -a + rt$ or $W_t = b + rt$, and let $p(a,b,r)$ be the probability that the exit is through the upper barrier—that $W_\tau = b + r\tau$.

(a) For the martingale $Y_{\theta,t}$ in the proof of Lemma 2, show that $E[Y_{\theta,t}] = 1$. Operating formally at first, conclude that

(37.48) $$E[e^{\theta W_\tau - \theta^2 \tau/2}] = 1.$$

Take $\theta = 2r$, and note that $\theta W_\tau - \frac{1}{2}\theta^2\tau$ is then $2rb$ if the exit is above (probability $p(a,b,r)$) and $-2ra$ if the exit is below (probability $1 - p(a,b,r)$). Deduce

$$p(a,b,r) = \frac{1 - e^{2ra}}{e^{2rb} - e^{-2ra}}.$$

(b) Show that $p(a,b,r) \to a/(a+b)$ as $r \to 0$, in agreement with (37.39).

(c) It remains to justify (37.48) for $\theta = 2r$. From $E[Y_{\theta,t}] = 1$ deduce

(37.49) $$E[e^{2r(W_\sigma - r^2\sigma)}] = 1$$

for nonrandom σ. By the arguments in the proofs of Lemmas 2 and 3, show that (37.49) holds for simple stopping times σ, for bounded ones, for $\sigma = \tau \wedge n$, for $\sigma = \tau$.

SECTION 38. NONDENUMERABLE PROBABILITIES*

Introduction

As observed a number of times above, the finite-dimensional distributions do not suffice to determine the character of the sample paths of a process. To obtain paths with natural regularity properties, the Poisson and Brownian motion processes were constructed by ad hoc methods. It is always possible to ensure that the paths have a certain very general regularity property called *separability*, and from this property will follow in appropriate circumstances various other desirable regularity properties.

Section 4 dealt with "denumerable" probabilities; questions about path functions involve all the time points and hence concern "nondenumerable" probabilities.

Example 38.1. For a mathematically simple illustration of the fact that path properties are not entirely determined by the finite-dimensional distributions, consider a probability space $(\Omega, \mathscr{F}, P)$ on which is defined a positive random variable V with continuous distribution: $P[V = x] = 0$ for each x. For $t \geq 0$, put $X(t, \omega) = 0$ for all ω, and put

$$(38.1) \qquad Y(t, \omega) = \begin{cases} 1 & \text{if } V(\omega) = t, \\ 0 & \text{if } V(\omega) \neq t. \end{cases}$$

Since V has continuous distribution, $P[X_t = Y_t] = 1$ for each t, and so $[X_t: t \geq 0]$ and $[Y_t: t \geq 0]$ are stochastic processes with identical finite-dimensional distributions; for each $t_1, \ldots, t_k$, the distribution $\mu_{t_1 \cdots t_k}$ common to $(X_{t_1}, \ldots, X_{t_k})$ and $(Y_{t_1}, \ldots, Y_{t_k})$ concentrates all its mass at the origin of R^k. But what about the sample paths? Of course, $X(\cdot, \omega)$ is identically 0, but $Y(\cdot, \omega)$ has a discontinuity—it is 1 at $t = V(\omega)$ and 0 elsewhere. It is because the position of this discontinuity has a continuous distribution that the two processes have the same finite-dimensional distributions. ∎

Definitions

The idea of separability is to make a countable set of time points serve to determine the properties of the process. In all that follows, the time set T will for definiteness be taken as $[0, \infty)$. Most of the results hold with an arbitrary subset of the line in the role of T.

As in Section 36, let R^T be the set of all real functions over $T = [0, \infty)$. Let D be a countable, dense subset of T. A function x—an element of R^T—is *separable D*, or *separable with respect to D*, if for each t in T there exists a

sequence $t_1, t_2, \ldots$ of points such that

$$(38.2) \qquad t_n \in D, \qquad t_n \to t, \qquad x(t_n) \to x(t).$$

(Because of the middle condition here, it was redundant to require D dense at the outset.) For t in D, (38.2) imposes no condition on x, since t_n may be taken as t. An x separable with respect to D is determined by its values at the points of D. Note, however, that separability requires that (38.2) hold for every t—an uncountable set of conditions. It is not hard to show that the set of functions separable with respect to D lies outside $\mathscr{R}^T$.

Example 38.2. If x is everywhere continuous or right-continuous, then it is separable with respect to every countable, dense D.

Suppose that $x(t)$ is 0 for $t \neq v$ and 1 for $t = v$, where $v > 0$. Then x is not separable with respect to D unless v lies in D. The paths $Y(\cdot, \omega)$ in Example 38.1 are of this form. ∎

The condition for separability can be stated another way: x is separable D if and only if for every t and every open interval I containing t, $x(t)$ lies in the closure of $[x(s): s \in I \cap D]$.

Suppose that x is separable D and that I is an open interval in T. If $\epsilon > 0$, then $x(t_0) + \epsilon > \sup_{t \in I} x(t) = u$ for some t_0 in I. By separability $|x(s_0) - x(t_0)| < \epsilon$ for some s_0 in $I \cap D$. But then $x(s_0) + 2\epsilon > u$, so that

$$(38.3) \qquad \sup_{t \in I} x(t) = \sup_{t \in I \cap D} x(t).$$

Similarly,

$$(38.4) \qquad \inf_{t \in I} x(t) = \inf_{t \in I \cap D} x(t)$$

and

$$(38.5) \qquad \sup_{t_0 \le t < t_0 + \delta} |x(t) - x(t_0)| = \sup_{\substack{t_0 \le t < t_0 + \delta \\ t \in D}} |x(t) - x(t_0)|.$$

A *stochastic process* $[X_t: t \ge 0]$ on $(\Omega, \mathscr{F}, P)$ is *separable D* if D is a countable, dense subset of $T = [0, \infty)$ and there is an $\mathscr{F}$-set N such that $P(N) = 0$ and such that the sample path $X(\cdot, \omega)$ is separable with respect to D for ω outside N. Finally, the process is separable if it is separable with respect to some D; this D is sometimes called a *separant*. In these definitions it is assumed for the moment that $X(t, \omega)$ is a *finite* real number for each t and ω.

Example 38.3. If the sample path $X(\cdot, \omega)$ is continuous for each ω, then the process is separable with respect to each countable, dense D. This covers Brownian motion as constructed in the preceding section. ∎

Example 38.4. Suppose that $[W_t: t \geq 0]$ has the finite-dimensional distributions of Brownian motion, but do not assume as in the preceding section that the paths are necessarily continuous. Assume, however, that $[W_t: t \geq 0]$ is separable with respect to D. Fix t_0 and δ. Choose sets $D_m = \{t_{m1}, \ldots, t_{mm}\}$ of D-points such that $t_0 < t_{m1} < \cdots < t_{mm} < t_0 + \delta$ and $D_m \uparrow D \cap (t_0, t_0 + \delta)$. By the argument leading to (37.9),

$$(38.6) \qquad P\left[\sup_{\substack{t_0 \leq t \leq t_0+\delta \\ t \in D}} |W_t - W_{t_0}| > \alpha \right] \leq \frac{K\delta^2}{\alpha^4}.$$

For sample points outside the N in the definition of separability, the supremum in (38.6) is unaltered, because of (38.5), if the restriction $t \in D$ is dropped. Since $P(N) = 0$,

$$P\left[\sup_{t_0 \leq t \leq t_0+\delta} |W_t - W_{t_0}| > \alpha \right] \leq \frac{K\delta^2}{\alpha^4}.$$

Define M_n by (37.8) but with r ranging over all the reals (not just over the dyadic rationals) in $[k2^{-n}, (k+2)2^{-n}]$. Then $P[M_n > n^{-i}] \leq 4Kn^5/2^n$ follows just as before. But for ω outside $B = [M_n > n^{-1} \text{ i.o.}]$, $W(\cdot, \omega)$ is continuous. Since $P(B) = 0$, $W(\cdot, \omega)$ is continuous for ω outside an $\mathscr{F}$-set of probability 0. If $(\Omega, \mathscr{F}, P)$ is complete, then the set of ω for which $W(\cdot, \omega)$ is continuous is an $\mathscr{F}$-set of probability 1. Thus paths are continuous with probability 1 for any separable process having the finite-dimensional distributions of Brownian motion—provided that the underlying space is complete, which can of course always be arranged.   ∎

As it will be shown below that there exists a separable process with any consistently prescribed set of finite-dimensional distributions, Example 38.4 provides another approach to the construction of continuous Brownian motion. The value of the method lies in its generality. It must not, however, be imagined that separability automatically ensures smooth sample paths:

Example 38.5. Suppose that the random variables X_t, $t \geq 0$, are independent, each having the standard normal distribution. Let D be any countable set dense in $T = [0, \infty)$. Suppose that I and J are open intervals with rational endpoints. Since the random variables X_t with $t \in D \cap I$ are independent, and since the value common to the $P[X_t \in J]$ is positive, the second Borel–Cantelli lemma implies that with probability 1, $X_t \in J$ for some t in $D \cap I$. Since there are only countably many pairs I and J with rational endpoints, there is an $\mathscr{F}$-set N such that $P(N) = 0$ and such that for ω outside N the set $[X(t, \omega): t \in D \cap I]$ is everywhere dense on the line for every open interval I in T. This implies that $[X_t: t \geq 0]$ is separable with respect to D. But also of course it implies that the paths are highly irregular.

This irregularity is not a shortcoming of the concept of separability—it is a necessary consequence of the properties of the finite-dimensional distributions specified in this example. ∎

Example 38.6. The process $[Y_t: t \geq 0]$ in Example 38.1 is not separable: The path $Y(\cdot, \omega)$ is not separable D unless D contains the point $V(\omega)$. The set of ω for which $Y(\cdot, \omega)$ is separable D is thus contained in $[\omega: V(\omega) \in D]$, a set of probability 0, since D is countable and V has a continuous distribution. ∎

Existence Theorems

It will be proved in stages that for every consistent system of finite-dimensional distributions there exists a separable process having those distributions. Define x to be separable D *at the point* t if there exist points t_n in D such that $t_n \to t$ and $x(t_n) \to x(t)$. Note that this is no restriction on x if t lies in D, and note that separability is the same thing as separability at every t.

Lemma 1. *Let* $[X_t: t \geq 0]$ *be a stochastic process on* $(\Omega, \mathscr{F}, P)$. *There exists a countable, dense set* D *in* $[0, \infty)$, *and there exists for each t an $\mathscr{F}$-set* $N(t)$, *such that* $P(N(t)) = 0$ *and such that for ω outside $N(t)$ the path function* $X(\cdot, \omega)$ *is separable D at t.*

PROOF. Fix open intervals I and J, and consider the probability

$$p(U) = P\left(\bigcap_{s \in U} [X_s \notin J] \right)$$

for countable subsets U of $I \cap T$. As U increases, the intersection here decreases and so does $p(U)$. Choose U_n so that $p(U_n) \to \inf_U p(U)$. If $U(I, J) = \bigcup_n U_n$, then $U(I, J)$ is a countable subset of $I \cap T$ making $p(U)$ minimal:

$$(38.7) \qquad P\left(\bigcap_{s \in U(I,J)} [X_s \notin J] \right) \leq P\left(\bigcap_{s \in U} [X_s \notin J] \right)$$

for every countable subset U of $I \cap T$. If $t \in I \cap T$, then

$$(38.8) \qquad P\left([X_t \in J] \cap \bigcap_{s \in U(I,J)} [X_s \notin J] \right) = 0,$$

because otherwise (38.7) would fail for $U = U(I, J) \cup \{t\}$.

Let $D = \bigcup U(I, J)$, where the union extends over all open intervals I and J with rational endpoints. Then D is a countable, dense subset of T. For each t let

$$(38.9) \qquad N(t) = \bigcup \left([X_t \in J] \cap \bigcap_{s \in U(I, J)} [X_s \notin J] \right),$$

where the union extends over all open intervals J that have rational endpoints and over all open intervals I that have rational endpoints and contain t. Then $N(t)$ is by (38.8) an $\mathscr{F}$-set such that $P(N(t)) = 0$.

Fix t and $\omega \notin N(t)$. The problem is to show that $X(\cdot, \omega)$ is separable with respect to D at t. Given n, choose open intervals I and J that have rational endpoints and lengths less than n^{-1} and satisfy $t \in I$ and $X(t, \omega) \in J$. Since ω lies outside (38.9), there must be an s_n in $U(I, J)$ such that $X(s_n, \omega) \in J$. But then $s_n \in D$, $|s_n - t| < n^{-1}$, and $|X(s_n, \omega) - X(t, \omega)| < n^{-1}$. Thus $s_n \to t$ and $X(s_n, \omega) \to X(t, \omega)$ for a sequence $s_1, s_2, \ldots$ in D. ∎

For any countable D, the set of ω for which $X(\cdot, \omega)$ is separable with respect to D at t is

$$(38.10) \qquad \bigcap_{n=1}^{\infty} \bigcup_{\substack{|s-t| < n^{-1} \\ s \in D}} \left[\omega : |X(t, \omega) - X(s, \omega)| < n^{-1} \right].$$

This set lies in $\mathscr{F}$ for each t, and the point of the lemma is that it is possible to choose D in such a way that each of these sets has probability 1.

Lemma 2. *Let $[X_t: t \geq 0]$ be a stochastic process on $(\Omega, \mathscr{F}, P)$. Suppose that for all t and ω*

$$(38.11) \qquad a < X(t, \omega) < b.$$

Then there exists on $(\Omega, \mathscr{F}, P)$ a process $[X_t': t \geq 0]$ having these three properties:

(i) *$P[X_t' = X_t] = 1$ for each t.*

(ii) *For some countable, dense subset D of $[0, \infty)$, $X'(\cdot, \omega)$ is separable D for every ω in Ω.*

(iii) *For all t and ω,*

$$(38.12) \qquad a \leq X'(t, \omega) \leq b.$$

PROOF. Choose a countable, dense set D and $\mathscr{F}$-sets $N(t)$ of probability 0 as in Lemma 1. If $t \in D$ or if $\omega \notin N(t)$, define $X'(t, \omega) = X(t, \omega)$. If $t \notin D$,

fix some sequence $\{s_n^{(t)}\}$ in D for which $\lim_n s_n^{(t)} = t$, and define $X'(t, \omega) = \limsup_n X(s_n^{(t)}, \omega)$ for $\omega \in N(t)$. To sum up,

$$(38.13) \quad X'(t, \omega) = \begin{cases} X(t, \omega) & \text{if } t \in D \text{ or } \omega \notin N(t), \\ \limsup_n X(s_n^{(t)}, \omega) & \text{if } t \notin D \text{ and } \omega \in N(t). \end{cases}$$

Since $N(t) \in \mathcal{F}$, X'_t is measurable $\mathcal{F}$ for each t. Since $P(N(t)) = 0$, $P[X_t = X'_t] = 1$ for each t.

Fix t and ω. If $t \in D$, then certainly $X'(\cdot, \omega)$ is separable D at t, and so assume $t \notin D$. If $\omega \notin N(t)$, then by the construction of $N(t)$, $X(\cdot, \omega)$ is separable with respect to D at t, so that there exist points s_n in D such that $s_n \to t$ and $X(s_n, \omega) \to X(t, \omega)$. But $X(s_n, \omega) = X'(s_n, \omega)$ because $s_n \in D$, and $X(t, \omega) = X'(t, \omega)$ because $\omega \notin N(t)$. Hence $X'(s_n, \omega) \to X'(t, \omega)$, and so $X'(\cdot, \omega)$ is separable with respect to D at t. Finally, suppose that $t \notin D$ and $\omega \in N(t)$. Then $X'(t, \omega) = \lim_k X(s_{n_k}^{(t)}, \omega)$ for some sequence $\{n_k\}$ of integers. As $k \to \infty$, $s_{n_k}^{(t)} \to t$ and $X'(s_{n_k}^{(t)}, \omega) = X(s_{n_k}^{(t)}, \omega) \to X'(t, \omega)$, so that again $X'(\cdot, \omega)$ is separable with respect to D at t. Clearly, (38.11) implies (38.12). ∎

Example 38.7. One must allow for the possibility of equality in (38.12). Suppose that $V(\omega) > 0$ for all ω and that V has a continuous distribution. Define

$$f(t) = \begin{cases} e^{-|t|} & \text{if } t \neq 0, \\ 0 & \text{if } t = 0, \end{cases}$$

and put $X(t, \omega) = f(t - V(\omega))$. If $[X'_t: t \geq 0]$ is any separable process with the same finite-dimensional distributions as $[X_t: t \geq 0]$, then $X'(\cdot, \omega)$ must with probability 1 assume the value 1 somewhere. In this case (38.11) holds for $a < 0$ and $b = 1$, and equality in (38.12) cannot be avoided. ∎

If

$$(38.14) \qquad\qquad \sup_{t, \omega} |X(t, \omega)| < \infty,$$

then (38.11) holds for some a and b. To treat the case in which (38.14) fails, it is necessary to allow for the possibility of infinite values. If $x(t)$ is ∞ or $-\infty$, replace the third condition in (38.2) by $x(t_n) \to \infty$ or $x(t_n) \to -\infty$. This extends the definition of separability to functions x that may assume infinite values and to processes $[X_t: t \geq 0]$ for which $X(t, \omega) = \pm\infty$ is a possibility.

Theorem 38.1. *If $[X_t: t \geq 0]$ is a finite-valued process on $(\Omega, \mathcal{F}, P)$, there exists on the same space a separable process $[X'_t: t \geq 0]$ such that $P[X'_t = X_t] = 1$ for each t.*

It is assumed for convenience here that $X(t, \omega)$ is finite for all t and ω, although this is not really necessary. But in some cases infinite values for certain $X'(t, \omega)$ cannot be avoided—see Example 38.8.

PROOF. If (38.14) holds, the result is an immediate consequence of Lemma 2. The definition of separability allows an exceptional set N of probability 0; in the construction of Lemma 2 this set is actually empty, but it is clear from the definition this could be arranged anyway.

The case in which (38.14) may fail could be treated by tracing through the preceding proofs, making slight changes to allow for infinite values. A simple argument makes this unnecessary. Let g be a continuous, strictly increasing mapping of R^1 onto $(0, 1)$. Let $Y(t, \omega) = g(X(t, \omega))$. Lemma 2 applies to $[Y_t: t \geq 0]$; there exists a separable process $[Y'_t: t \geq 0]$ such that $P[Y'_t = Y_t] = 1$. Since $0 < Y(t, \omega) < 1$, Lemma 2 ensures $0 \leq Y'(t, \omega) \leq 1$. Define

$$X'(t, \omega) = \begin{cases} -\infty & \text{if } Y'(t, \omega) = 0, \\ g^{-1}(Y'(t, \omega)) & \text{if } 0 < Y'(t, \omega) < 1, \\ +\infty & \text{if } Y'(t, \omega) = 1. \end{cases}$$

Then $[X'_t: t \geq 0]$ satisfies the requirements. Note that $P[X'_t = \pm\infty] = 0$ for each t.   ∎

Example 38.8. Suppose that $V(\omega) > 0$ for all ω and V has a continuous distribution. Define

$$h(t) = \begin{cases} |t|^{-1} & \text{if } t \neq 0, \\ 0 & \text{if } t = 0, \end{cases}$$

and put $X(t, \omega) = h(t - V(\omega))$. This is analogous to Example 38.7. If $[X'_t: t \geq 0]$ is separable and has the finite-dimensional distributions of $[X_t: t \geq 0]$, then $X'(\cdot, \omega)$ must with probability 1 assume the value ∞ for some t.   ∎

Combining Theorem 38.1 with Kolmogorov's existence theorem shows that *for any consistent system of finite-dimensional distributions $\mu_{t_1 \cdots t_k}$ there exists a separable process with the $\mu_{t_1 \cdots t_k}$ as finite-dimensional distributions.* As shown in Example 38.4, this leads to another construction of Brownian motion with continuous paths.

Consequences of Separability

The next theorem implies in effect that, if the finite-dimensional distributions of a process are such that it "should" have continuous paths, then it will in fact have continuous paths if it is separable. Example 38.4 illustrates this. The same thing holds for properties other than continuity.

Let $\overline{R}^T$ be the set of functions on $T = [0, \infty)$ with values that are ordinary reals or else ∞ or $-\infty$. Thus $\overline{R}^T$ is an enlargement of the R^T of Section 36, an enlargement necessary because separability sometimes forces infinite values. Define the function Z_t on $\overline{R}^T$ by $Z_t(x) = Z(t, x) = x(t)$. This is just an extension of the coordinate function (36.8). Let $\overline{\mathscr{R}}^T$ be the σ-field in $\overline{R}^T$ generated by the Z_t, $t \geq 0$.

Suppose that A is a subset of $\overline{R}^T$, not necessarily in $\overline{\mathscr{R}}^T$. For $D \subset T = [0, \infty)$, let A_D consist of those elements x of $\overline{R}^T$ that agree on D with some element y of A:

$$(38.15) \qquad A_D = \bigcup_{y \in A} \bigcap_{t \in D} \left[x \in \overline{R}^T : x(t) = y(t) \right].$$

Of course, $A \subset A_D$. Let S_D denote the set of x in $\overline{R}^T$ that are separable with respect to D.

In the following theorem, $[X_t: t \geq 0]$ and $[X_t': t \geq 0]$ are processes on spaces $(\Omega, \mathscr{F}, P)$ and $(\Omega', \mathscr{F}', P')$, which may be distinct; the path functions are $X(\cdot, \omega)$ and $X'(\cdot, \omega')$.

Theorem 38.2. *Suppose of A that for each countable, dense subset D of $T = [0, \infty)$, the set (38.15) satisfies*

$$(38.16) \qquad A_D \in \overline{\mathscr{R}}^T, \qquad A_D \cap S_D \subset A.$$

If $[X_t: t \geq 0]$ and $[X_t': t \geq 0]$ have the same finite-dimensional distributions, if $[\omega: X(\cdot, \omega) \in A]$ lies in $\mathscr{F}$ and has P-measure 1, and if $[X_t': t \geq 0]$ is separable, then $[\omega': X'(\cdot, \omega') \in A]$ contains an $\mathscr{F}'$-set of P'-measure 1.

If $(\Omega', \mathscr{F}', P')$ is complete, then of course $[\omega': X'(\cdot, \omega') \in A]$ is itself an $\mathscr{F}'$-set of P'-measure 1.

PROOF. Suppose that $[X_t': t \geq 0]$ is separable with respect to D. The difference $[\omega': X'(\cdot, \omega') \in A_D] - [\omega': X'(\cdot, \omega') \in A]$ is by (38.16) a subset of $[\omega': X'(\cdot, \omega') \in \overline{R}^T - S_D]$, which is contained in an $\mathscr{F}'$-set of N' of P'-measure 0. Since the two processes have the same finite-dimensional distributions and hence induce the same distribution on $(\overline{R}^T, \overline{\mathscr{R}}^T)$, and since A_D lies in $\overline{\mathscr{R}}^T$, it follows that $P'[\omega': X'(\cdot, \omega') \in A_D] = P[\omega: X(\cdot, \omega) \in A_D] \geq P[\omega: X(\cdot, \omega) \in A] = 1$. Thus the subset $[\omega': X'(\cdot, \omega') \in A_D] - N'$ of $[\omega': X'(\cdot, \omega') \in A]$ lies in $\mathscr{F}'$ and has P'-measure 1. ∎

Example 38.9. Consider the set C of finite-valued, continuous functions on T. If $x \in S_D$ and $y \in C$, and if x and y agree on a dense D, then x and y agree everywhere: $x = y$. Therefore, $C_D \cap S_D \subset C$. Further,

$$C_D = \bigcap_{\epsilon, t} \bigcup_{\delta} \bigcap_{s} \left[x \in \overline{R}^T : |x(s)| < \infty, |x(t)| < \infty, |x(s) - x(t)| < \epsilon \right],$$

where ϵ and δ range over the positive rationals, t ranges over D, and the inner intersection extends over the s in D satisfying $|s - t| < \delta$. Hence $C_D \in \mathscr{R}^T$. Thus C satisfies the condition (38.16).

Theorem 38.2 now implies that if a process has continuous paths with probability 1, then any separable process having the same finite-dimensional distributions has continuous paths outside a set of probability 0. In particular, a Brownian motion with continuous paths was constructed in the preceding section, and so any separable process with the finite-dimensional distributions of Brownian motion has continuous paths outside a set of probability 0. The argument in Example 38.4 now becomes supererogatory.   ∎

Example 38.10. There is a somewhat similar argument for the step functions of the Poisson process. Let Z^+ be the set of nonnegative integers; let E consist of the nondecreasing functions x in $\overline{R}^T$ such that $x(t) \in Z^+$ for all t and such that for every $n \in Z^+$ there exists a nonempty interval I such that $x(t) = n$ for $t \in I$. Then

$$E_D = \bigcap_{t \in D} [x : x(t) \in Z^+] \cap \bigcap_{s, t \in D, s < t} [x : x(s) \leq x(t)]$$

$$\cap \bigcap_{n=0}^{\infty} \bigcup_I \bigcap_{t \in D \cap I} [x : x(t) = n],$$

where I ranges over the open intervals with rational endpoints. Thus $E_D \in \mathscr{R}^T$. Clearly, $E_D \cap S_D \subset E$, and so Theorem 38.2 applies.

In Section 23 was constructed a Poisson process with paths in E, and therefore any separable process with the same finite-dimensional distributions will have paths in E except for a set of probability 0.   ∎

Example 38.11. For E as in Example 38.10, let E_0 consist of the elements of E that are right-continuous; a function in E need not lie in E_0, although at each t it must be continuous from one side or the other. The Poisson process as defined in Section 23 by $N_t = \max[n : S_n \leq t]$ (see (23.5)) has paths in E_0. But if $N_t' = \max[n : S_n < t]$, then $[N_t' : t \geq 0]$ is separable and has the same finite-dimensional distributions, but its paths are not in E_0. Thus E_0 does not satisfy the hypotheses of Theorem 38.2. Separability does not help distinguish between continuity from the right and continuity from the left.   ∎

Example 38.12. The class of sets A satisfying (38.16) is closed under the formation of countable unions and intersections but is not closed under complementation. Define X_t and Y_t as in Example 38.1, and let C be the set of continuous paths. Then $[Y_t : t \geq 0]$ and $[X_t : t \geq 0]$ have the same finite-dimensional distributions, and the latter is separable; $Y(\cdot, \omega)$ is in $\overline{R}^T - C$ for each ω, and $X(\cdot, \omega)$ is in $\overline{R}^T - C$ for no ω.   ∎

Example 38.13. As a final example, consider the set J of functions with discontinuities of at most the first kind: x is in J if it is finite-valued, if $x(t +) = \lim_{s \downarrow t} x(s)$ exists (finite) for $t \geq 0$ and $x(t -) = \lim_{s \uparrow t} x(s)$ exists (finite) for $t > 0$, and if $x(t)$ lies between $x(t +)$ and $x(t -)$ for $t > 0$. Continuous and right-continuous functions are special cases.

Let V denote the general system

$$(38.17) \qquad V: k; r_1, \ldots, r_k; s_1, \ldots, s_k; \alpha_1, \ldots, \alpha_k,$$

where k is an integer, where the r_i, s_i, and α_i are rational, and where

$$0 = r_1 < s_1 < r_2 < s_2 < \cdots < r_k < s_k.$$

Define

$$J(D, V, \epsilon) = \bigcap_{i=1}^{k} \left[x : \alpha_i \leq x(t) \leq \alpha_i + \epsilon, t \in (r_i, s_i) \cap D \right]$$

$$\cap \bigcap_{i=2}^{k} \left[x : \min\{\alpha_{i-1}, \alpha_i\} \leq x(t) \leq \max\{\alpha_{i-1}, \alpha_i\} + \epsilon, t \in (s_{i-1}, r_i) \cap D \right]$$

Let $\mathcal{U}_{m,k,\delta}$ be the class of systems (38.17) that have a fixed value for k and satisfy $r_i - s_{i-1} < \delta$, $i = 2, \ldots, k$, and $s_k > m$. It will be shown that

$$(38.18) \qquad J_D = \bigcap_{m=1}^{\infty} \bigcap_{\epsilon} \bigcup_{k=1}^{\infty} \bigcap_{\delta} \bigcup_{V \in \mathcal{U}_{m,k,\delta}} J(D, V, \epsilon),$$

where ϵ and δ range over the positive rationals. From this it will follow that $J_D \in \mathscr{R}^T$. It will also be shown that $J_D \cap S_D \subset J$, so that J satisfies the hypothesis of Theorem 38.2.

Suppose that $y \in J$. For fixed ϵ, let H be the set of nonnegative h for which there exist finitely many points t_i such that $0 = t_0 \leq t_1 \leq \cdots \leq t_r = h$ and $|y(t) - y(t')| < \epsilon$ for t and t' in the same interval (t_{i-1}, t_i). If $h_n \in H$ and $h_n \uparrow h$, then from the existence of $y(h-)$ follows $h \in H$. Hence H is closed. If $h \in H$, from the existence of $y(h+)$ it follows that H contains points to the right of h. Therefore, $H = [0, \infty)$. From this it follows that the right side of (38.18) contains J_D.

Suppose that x is a member of the right side of (38.18). It is not hard to deduce that for each t the limits

$$(38.19) \qquad \lim_{s \downarrow t, s \in D} x(s), \qquad \lim_{s \uparrow t, s \in D} x(s)$$

exist and that $x(t)$ lies between them if $t \in D$. For $t \in D$ take $y(t) = x(t)$, and for $t \notin D$ take $y(t)$ to be the first limit in (38.19). Then $y \in J$ and hence $x \in J_D$. This argument also shows that $J_D \cap S_D \subset J$. ∎

Appendix

Gathered here for easy reference are certain definitions and results from set theory and real analysis required in the text. Although there are many newer books, HAUSDORFF (the early sections) on set theory and HARDY on analysis are still excellent for the general background assumed here.

Set Theory

A1. The *empty set* is denoted by $\varnothing$. *Sets* are variable subsets of some *space* that is fixed in any one definition, argument, or discussion; this space is denoted either generically by Ω or by some special symbol (such as R^k for Euclidean k-space). A *singleton* is a set consisting of just one point or element. That A is a *subset* of B is expressed by $A \subset B$. In accordance with standard usage, $A \subset B$ does not preclude $A = B$; A is a *proper* subset of B if $A \subset B$ and $A \neq B$.

The *complement* of A is always relative to the overall space Ω; it consists of the points of Ω not contained in A and is denoted by A^c. The *difference* between A and B, denoted by $A - B$, is $A \cap B^c$; here B need not be contained in A, and if it is, then $A - B$ is a *proper* difference. The *symmetric difference* $A \vartriangle B = (A \cap B^c) \cup (A^c \cap B)$ consists of the points that lie in one of the sets A and B but not in both.

Classes of sets are denoted by script letters. The *power set* of Ω is the class of all subsets of Ω; it is denoted 2^{Ω}.

A2. The set of ω that lie in A and satisfy a given property $p(\omega)$ is denoted $[\omega \in A: p(\omega)]$; if $A = \Omega$, this is usually shortened to $[\omega: p(\omega)]$.

A3. In this book, to say that a collection $[A_\theta: \theta \in \Theta]$ is *disjoint* always means that it is *pairwise* disjoint: $A_\theta \cap A_{\theta'} = 0$ if θ and θ' are distinct elements of the index set Θ. To say that A *meets* B, or that B meets A, is to say that they are not disjoint: $A \cap B \neq \varnothing$. The collection $[A_\theta: \theta \in \Theta]$ *covers* B if $B \subset \bigcup_\theta A_\theta$. The collection is a *decomposition* or *partition* of B if it is disjoint and $B = \bigcup_\theta A_\theta$.

A4. By $A_n \uparrow A$ is meant $A_1 \subset A_2 \subset \cdots$ and $A = \bigcup_n A_n$; by $A_n \downarrow A$ is meant $A_1 \supset A_2 \supset \cdots$ and $A = \bigcap_n A_n$.

A5. The *indicator*, or *indicator function*, of a set A is the function on Ω that assumes the value 1 on A and 0 on A^c; it is denoted I_A. The alternative term "characteristic function" is reserved for the Fourier transform (see Section 26).

536

A6. *De Morgan's* laws are $(\bigcup_\theta A_\theta)^c = \bigcap_\theta A_\theta^c$ and $(\bigcap_\theta A_\theta)^c = \bigcup_\theta A_\theta^c$. These and the other facts of basic set theory are assumed known: a countable union of countable sets is countable, and so on.

A7. If $T: \Omega \to \Omega'$ is a mapping of Ω into Ω' and A' is a set in Ω', the *inverse image* of A' is $T^{-1}A' = [\omega \in \Omega: T\omega \in A']$. It is easily checked that each of these statements is equivalent to the next: $\omega \in \Omega - T^{-1}A'$, $\omega \notin T^{-1}A'$, $T\omega \notin A'$, $T\omega \in \Omega' - A'$, $\omega \in T^{-1}(\Omega' - A')$. Therefore, $\Omega - T^{-1}A' = T^{-1}(\Omega' - A')$. Simple considerations of this kind show that $\bigcup_\theta T^{-1}A'_\theta = T^{-1}(\bigcup_\theta A'_\theta)$ and $\bigcap_\theta T^{-1}A'_\theta = T^{-1}(\bigcap_\theta A'_\theta)$, and that $A' \cap B' = \varnothing$ implies $T^{-1}A' \cap T^{-1}B' = \varnothing$ (the reverse implication is false unless $T\Omega = \Omega'$).

If f maps Ω into another space, $f(\omega)$ is the value of the function f at an unspecified value of the argument ω. The function f itself (the rule defining the mapping) is sometimes denoted $f(\cdot)$. This is especially convenient for a function $f(\omega, t)$ of two arguments: For each fixed t, $f(\cdot, t)$ denotes the function on Ω with value $f(\omega, t)$ at ω.

A8. *The axiom of choice.* Suppose that $[A_\theta: \theta \in \Theta]$ is a decomposition of Ω into nonempty sets. The axiom of choice says that there exists a set (at least one set) C that contains exactly one point from each A_θ: $C \cap A_\theta$ is a singleton for each θ in Θ. The existence of such sets C is assumed in "everyday" mathematics, and the axiom of choice may even seem to be simply *true*. A careful treatment of set theory, however, is based on an explicit list of such axioms and a study of the relationships between them; see HALMOS₂ or DUDLEY.

A few of the problems require *Zorn's lemma*, which is equivalent to the axiom of choice; see DUDLEY or KAPLANSKY.

The Real Line

A9. The real line is denoted by R^1; $x \vee y = \max\{x, y\}$ and $x \wedge y = \min\{x, y\}$. For real x, $\lfloor x \rfloor$ is the integer part of x, and sgn x is $+1$, 0, or -1 as x is positive, 0, or negative. It is convenient to be explicit about open, closed, and half-open intervals:

$$(a, b) = [x: a < x < b],$$

$$[a, b] = [x: a \le x \le b],$$

$$(a, b] = [x: a < x \le b],$$

$$[a, b) = [x: a \le x < b].$$

A10. Of course $x_n \to x$ means $\lim_n x_n = x$; $x_n \uparrow x$ means $x_1 \le x_2 \le \cdots$ and $x_n \to x$; $x_n \downarrow x$ means $x_1 \ge x_2 \ge \cdots$ and $x_n \to x$.

A sequence $\{x_n\}$ *is bounded if and only if every subsequence* $\{x_{n_k}\}$ *contains a further subsequence* $\{x_{n_{k(j)}}\}$ *that converges to some* x: $\lim_j x_{n_{k(j)}} = x$. If $\{x_n\}$ is not bounded, then for each k there is an n_k for which $|x_{n_k}| > k$; no subsequence of $\{x_{n_k}\}$ can converge. The implication in the other direction is a simple consequence of the fact that every bounded sequence contains a convergent subsequence.

If $\{x_n\}$ *is bounded, and if each subsequence that converges at all converges to* x, *then* $\lim_n x_n = x$. If x_n does not converge to x, then $|x_{n_k} - x| > \epsilon$ for some positive ϵ and some increasing sequence $\{n_k\}$ of integers; some subsequence of $\{x_{n_k}\}$ converges, but the limit cannot be x.

A11. A set G is defined as *open* if for each x in G there is an open interval I such that $x \in I \subset G$. A set F is defined as *closed* if F^c is open. The *interior* of A, denoted

A°, consists of the x in A for which there exists an open interval I such that $x \in I \subset A$. The *closure* of A, denoted A^-, consists of the x for which there exists a sequence $\{x_n\}$ in A with $x_n \to x$. The *boundary* of A is $\partial A = A^- - A^{\circ}$. The basic facts of real analysis are assumed known: A is open if and only if $A = A^{\circ}$; A is closed if and only if $A = A^-$; A is closed if and only if it contains all limits of sequences in it; x lies in ∂A if and only if there is a sequence $\{x_n\}$ in A and a sequence $\{y_n\}$ in A^c such that $x_n \to x$ and $y_n \to x$; and so on.

A12. Every open set G on the line is a countable, disjoint union of open intervals. To see this, define points x and y of G to be equivalent if $x \le y$ and $[x, y] \subset G$ or $y \le x$ and $[y, x] \subset G$. This is an equivalence relation. Each equivalence class is an interval, and since G is open, each is in fact an open interval. Thus G is a disjoint union of open (nonempty) intervals, and there can be only countably many of them, since each contains a rational.

A13. The simplest form of the *Heine-Borel* theorem says that if $[a, b] \subset \bigcup_{k=1}^{\infty}(a_k, b_k)$, then $[a, b] \subset \bigcup_{k=1}^{n}(a_k, b_k)$ for some n. A set A is defined to be *compact* if each cover of it by open sets has a finite subcover—that is, if $[G_\theta: \theta \in \Theta]$ covers A and each G_θ is open, then some finite subcollection $\{G_{\theta_1}, \ldots, G_{\theta_n}\}$ covers A. Equivalent to the Heine-Borel theorem is the assertion that a bounded, closed set is compact. Also equivalent is the assertion that every bounded sequence of real numbers has a convergent subsequence.

A14. *The diagonal method.* From this last fact follows one of the basic principles of analysis.

Theorem. *Suppose that each row of the array*

(1)
$$
\begin{array}{cccc}
x_{1,1} & x_{1,2} & x_{1,3} & \cdots \\
x_{2,1} & x_{2,2} & x_{2,3} & \cdots \\
\vdots & \vdots & \vdots &
\end{array}
$$

is a bounded sequence of real numbers. Then there exists an increasing sequence $n_1, n_2, \ldots$ of integers such that the limit $\lim_k x_{r,n_k}$ exists for $r = 1, 2, \ldots$.

PROOF. From the first row, select a convergent subsequence

(2)
$$
x_{1,n_{1,1}}, x_{1,n_{1,2}}, x_{1,n_{1,3}}, \ldots;
$$

here $\{n_{1,k}\}$ is an increasing sequence of integers and $\lim_k x_{1,n_{1,k}}$ exists. Look next at the second row of (1) along the sequence $n_{1,1}, n_{1,2}, \ldots$:

(3)
$$
x_{2,n_{1,1}}, x_{2,n_{1,2}}, x_{2,n_{1,3}}, \ldots.
$$

As a subsequence of the second row of (1), (3) is bounded. Select from it a convergent subsequence

$$
x_{2,n_{2,1}}, x_{2,n_{2,2}}, x_{2,n_{2,3}}, \ldots;
$$

here $\{n_{2,k}\}$ is an increasing sequence of integers, a subsequence of $\{n_{1,k}\}$, and $\lim_k x_{2,n_{2,k}}$ exists.

Continue inductively in the same way. This gives an array

(4)
$$
\begin{array}{cccc}
n_{1,1} & n_{1,2} & n_{1,3} & \cdots \\
n_{2,1} & n_{2,2} & n_{2,3} & \cdots \\
\vdots & \vdots & \vdots &
\end{array}
$$

with three properties: (i) Each row of (4) is an increasing sequence of integers. (ii) The rth row is a subsequence of the $(r-1)$st. (iii) For each r, $\lim_k x_{r,n_{r,k}}$ exists. Thus

(5)
$$
x_{r,n_{r,1}}, x_{r,n_{r,2}}, x_{r,n_{r,3}}, \ldots
$$

is a convergent subsequence of the rth row of (1).

Put $n_k = n_{k,k}$. Since each row of (4) is increasing and is contained in the preceding row, $n_1, n_2, n_3, \ldots$ is an increasing sequence of integers. Furthermore, $n_r, n_{r+1}, n_{r+2}, \ldots$ is a subsequence of the rth row of (4). Thus $x_{r,n_r}, x_{r,n_{r+1}}, x_{r,n_{r+2}}, \ldots$ is a subsequence of (5) and is therefore convergent. Thus $\lim_k x_{r,n_k}$ does exist. ∎

Since $\{n_k\}$ is the diagonal of the array (4), application of this theorem is called the *diagonal method*.

A15. The set A is by definition *dense in* the set B if for each x in B and each open interval J containing x, J meets A. This is the same thing as requiring $B \subset A^-$. The set E is by definition *nowhere dense* if each open interval I contains some open interval J that does not meet E. This makes sense: If I contains an open interval J that does not meet E, then E is not dense in I; the definition requires that E be dense in *no* interval I.

A set A is defined to be *perfect* if it is closed and for each x in A and positive ϵ, there is a y in A such that $0 < |x - y| < \epsilon$. An equivalent requirement is that A be closed and for each x in A there exist a sequence $\{x_n\}$ in A such that $x_n \neq x$ and $x_n \to x$. The Cantor set is uncountable, nowhere dense, and perfect.

A set that is nowhere dense is in a sense small. If A is a countable union of sets each of which is nowhere dense, then A is said to be of the *first category*. This is a weaker notion of smallness. A set that is not of the first category is said to be of the *second category*.

Euclidean k-Space

A16. Euclidean space of dimension k is denoted R^k. Points $(a_1, \ldots, a_k)$ and $(b_1, \ldots, b_k)$ determine open, closed, and half-open rectangles in R^k:

$$
[x: a_i < x_i < b_i, i = 1, \ldots, k],
$$
$$
[x: a_i \leq x_i \leq b_i, i = 1, \ldots, k],
$$
$$
[x: a_i < x_i \leq b_i, i = 1, \ldots, k].
$$

A rectangle (without a qualifying adjective) is in this book a set of this last form.

The Euclidean distance $(\sum_{i=1}^{k} (x_i - y_i)^2)^{1/2}$ between $x = (x_1, \ldots, x_k)$ and $y = (y_1, \ldots, y_k)$ is denoted by $|x - y|$.

A17. All the concepts in A11 carry over to R^k: simply take the I there to be an open rectangle in R^k. The definition of compact set also carries over word for word, and the Heine–Borel theorem in R^k says that a closed, bounded set is compact.

Analysis

A18. The standard Landau notation is used. Suppose that $\{x_n\}$ and $\{y_n\}$ are real sequences and $y_n > 0$. Then $x_n = O(y_n)$ means x_n/y_n is bounded; $x_n = o(y_n)$ means $x_n/y_n \to 0$; $x_n \sim y_n$ means $x_n/y_n \to 1$; $x_n \asymp y_n$ means x_n/y_n and y_n/x_n are both bounded. To write $x_n = z_n + O(y_n)$, for example, means that $x_n = z_n + u_n$ for some $\{u_n\}$ satisfying $u_n = O(y_n)$—that is, that $x_n - z_n = O(y_n)$.

A19. *A difference equation.* Suppose that a and b are integers and $a < b$. Suppose that x_n is defined for $a \le n \le b$ and satisfies

$$(6) \qquad x_n = px_{n+1} + qx_{n-1} \qquad \text{for } a < n < b,$$

where p and q are positive and $p + q = 1$. The general solution of this difference equation has the form

$$(7) \qquad x_n = \begin{cases} A + B(q/p)^n & \text{for } a \le n \le b \quad \text{if } p \ne q, \\ A + Bn & \text{for } a \le n \le b \quad \text{if } p = q. \end{cases}$$

That (7) always solves (6) is easily checked. Suppose the values x_{n_1} and x_{n_2} are given, where $a \le n_1 < n_2 \le b$. If $p \ne q$, the system

$$A + B(q/p)^{n_1} = x_{n_1}, \qquad A + B(q/p)^{n_2} = x_{n_2}$$

can always be solved for A and B. If $p = q$, the system

$$A + Bn_1 = x_{n_1}, \qquad A + Bn_2 = x_{n_2}$$

can always be solved. Take $n_1 = a$ and $n_2 = a + 1$; the corresponding A and B satisfy (7) for $n = a$ and for $n = a + 1$, and it follows by induction that (7) holds for all n. Thus any solution of (6) can indeed be put in the form (7). Furthermore, the equation (6) and any pair of values x_{n_1} and x_{n_2} ($n_1 \ne n_2$) suffice to determine all the x_n.

If x_n is defined for $a \le n < \infty$ and satisfies (6) for $a < n < \infty$, then there are constants A and B such that (7) holds for $a \le n < \infty$.

A20. *Cauchy's equation.*

Theorem. *Let f be a real function on $(0, \infty)$, and suppose that f satisfies Cauchy's equation: $f(x + y) = f(x) + f(y)$ for $x, y > 0$. If there is some interval on which f is bounded above, then $f(x) = xf(1)$ for $x > 0$.*

PROOF. The problem is to prove that $g(x) = f(x) - xf(1)$ vanishes identically. Clearly, $g(1) = 0$, and g satisfies Cauchy's equation and on some interval is bounded above. By induction, $g(nx) = ng(x)$; hence $ng(m/n) = g(m) = mg(1) = 0$, so that $g(r) = 0$ for positive rational r. Suppose that $g(x_0) \ne 0$ for some x_0. If $g(x_0) < 0$, then $g(r_0 - x_0) = -g(x_0) > 0$ for rational $r_0 > x_0$. It is thus no restriction to assume that $g(x_0) > 0$. Let I be an open interval in which g is bounded above. Given a number M, choose n so that $ng(x_0) > M$, and then choose a rational r so that $nx_0 + r$ lies in I. If $r > 0$, then $g(r + nx_0) = g(r) + g(nx_0) = g(nx_0) = ng(x_0)$. If $r < 0$, then $ng(x_0) = g(nx_0) = g((-r) + (nx_0 + r)) = g(-r) + g(nx_0 + r) = g(nx_0 + r)$. In either

case, $g(nx_0) + r) = ng(x_0)$; of course this is trivial if $r = 0$. Since $g(nx_0 + r) = ng(x_0) > M$ and M was arbitrary, g is not bounded above in I, a contradiction. ∎

Obviously, the same proof works if f is bounded below in some interval.

Corollary. Let U be a real function on $(0, \infty)$ and suppose that $U(x + y) = U(x)U(y)$ for $x, y > 0$. Suppose further that there is some interval on which U is bounded above. Then either $U(x) = 0$ for $x > 0$, or else there is an A such that $U(x) = e^{Ax}$ for $x > 0$.

PROOF. Since $U(x) = U^2(x/2)$, U is nonnegative. If $U(x) = 0$, then $U(x/2^n) = 0$ and so U vanishes at points arbitrarily near 0. If U vanishes at a point, it must by the functional equation vanish everywhere to the right of that point. Hence U is identically 0 or else everywhere positive.

In the latter case, the theorem applies to $f(x) = \log U(x)$, this function being bounded above in some interval, and so $f(x) = Ax$ for $A = \log U(1)$. ∎

A21. A number-theoretic fact.

Theorem. Suppose that M is a set of positive integers closed under addition and that M has greatest common divisor 1. Then M contains all integers exceeding some n_0.

PROOF. Let M_1 consist of all the integers m, $-m$, and $m - m'$ with m and m' in M. Then M_1 is closed under addition and subtraction (it is a subgroup of the group of integers). Let d be the smallest positive element of M_1. If $n \in M_1$, write $n = qd + r$, where $0 \le r < d$. Since $r = n - qd$ lies in M_1, r must actually be 0. Thus M_1 consists of the multiples of d. Since d divides all the integers in M_1 and hence all the integers in M, and since M has greatest common divisor 1, $d = 1$. Thus M_1 contains all the integers.

Write $1 = m - m'$ with m and m' in M (if 1 itself is in M, the proof is easy), and take $n_0 = (m + m')^2$. Given $n > n_0$, write $n = q(m + m') + r$, where $0 \le r < m + m'$. From $n > n_0 \ge (r + 1)(m + m')$ follows $q = (n - r)/(m + m') > r$. But $n = q(m + m') + r(m - m') = (q + r)m + (q - r)m'$, and since $q + r \ge q - r > 0$, n lies in M. ∎

A22. One- and two-sided derivatives.

Theorem. Suppose that f and g are continuous on $[0, \infty)$ and g is the right-hand derivative of f on $(0, \infty)$: $f^+(t) = g(t)$ for $t > 0$. Then $f^+(0) = g(0)$ as well, and g is the two-sided derivative of f on $(0, \infty)$.

PROOF. It suffices to show that $F(t) = f(t) - f(0) - \int_0^t g(s)\,ds$ vanishes for $t \ge 0$. By assumption, F is continuous on $[0, \infty)$ and $F^+(t) = 0$ for $t > 0$. Suppose that $F(t_0) > F(t_1)$, where $0 < t_0 < t_1$. Then $G(t) = F(t) - (t - t_0)(F(t_1) - F(t_0))/(t_1 - t_0)$ is continuous on $[0, \infty)$, $G(t_0) = G(t_1)$, and $G^+(t) > 0$ on $(0, \infty)$. But then the maximum of G over $[t_0, t_1]$ must occur at some interior point; since $G^+ \le 0$ at a local maximum, this is impossible. Similarly $F(t_0) < F(t_1)$ is impossible. Thus F is constant over $(0, \infty)$ and by continuity is constant over $[0, \infty)$. Since $F(0) = 0$, F vanishes on $[0, \infty)$. ∎

A23. A differential equation.
The equation $f'(t) = Af(t) + g(t)$ ($t \ge 0$; g continuous) has the particular solution $f_0(t) = e^{At}\int_0^t g(s)e^{-As}\,ds$; for an arbitrary solution f, $(f(t) - f_0(t))e^{-At}$ has derivative 0 and hence equals $f(0)$ identically. All solutions thus have the form $f(t) = e^{At}[f(0) + \int_0^t g(s)e^{-As}\,ds]$.

A24. *A trigonometric identity.* If $z \neq 1$ and $z \neq 0$, then

$$\sum_{k=-l}^{l} z^k = z^{-l} \sum_{k=0}^{2l} z^k = z^{-l} \frac{1 - z^{2l+1}}{1 - z} = \frac{z^{-l} - z^{l+1}}{1 - z},$$

and hence

$$\sum_{l=0}^{m-1} \sum_{k=-l}^{l} z^k = \sum_{l=0}^{m-1} \frac{z^{-l} - z^{l+1}}{1 - z} = \frac{1}{1 - z} \left[\frac{1 - z^{-m}}{1 - z^{-1}} - z \frac{1 - z^m}{1 - z} \right]$$

$$= \frac{1 - z^{-m} + 1 - z^m}{(1 - z)(1 - z^{-1})} = \frac{\left(z^{m/2} - z^{-m/2} \right)^2}{\left(z^{1/2} - z^{-1/2} \right)^2}.$$

Take $z = e^{ix}$. If x is not an integral multiple of 2π, then

$$\sum_{l=0}^{m-1} \sum_{k=-l}^{l} e^{ikx} = \frac{\left(\sin \frac{1}{2}mx \right)^2}{\left(\sin \frac{1}{2}x \right)^2}.$$

If $x = 2\pi n$, the left-hand side here is m^2, which is the limit of the right-hand side as $x \to 2\pi n$.

Infinite Series

A25. *Nonnegative series.* Suppose $x_1, x_2, \ldots$ are nonnegative. If E is a finite set of integers, then $E \subset \{1, 2, \ldots, n\}$ for some n, so that by nonnegativity $\sum_{k \in E} x_k \leq \sum_{k=1}^{n} x_k$. The set of partial sums $\sum_{k=1}^{n} x_k$ thus has the same supremum as the larger set of sums $\sum_{k \in E} x_k$ (E finite). Therefore, the nonnegative series $\sum_{k=1}^{\infty} x_k$ converges if and only if the sums $\sum_{k \in E} x_k$ for finite E are bounded, in which case the sum is the supremum: $\sum_{k=1}^{\infty} x_k = \sup_E \sum_{k \in E} x_k$.

A26. *Dirichlet's theorem.* Since the supremum in A25 is invariant under permutations, so is $\sum_{k=1}^{\infty} x_k$: If the x_k are nonnegative and $y_k = x_{f(k)}$ for some one-to-one map f of the positive integers onto themselves, then $\sum_k x_k$ and $\sum_k y_k$ diverge or converge together and in the latter case have the same sum.

A27. *Double series.* Suppose that x_{ij}, $i, j = 1, 2, \ldots$, are nonnegative. The ith row gives a series $\sum_j x_{ij}$, and if each of these converges, one can form the series $\sum_i \sum_j x_{ij}$. Let the terms x_{ij} be arranged in some order as a single infinite series $\sum_{ij} x_{ij}$; by Dirichlet's theorem, the sum is the same whatever order is used.

Suppose each $\sum_j x_{ij}$ converges and $\sum_i \sum_j x_{ij}$ converges. If E is a finite set of the pairs (i, j), there is an n for which $\sum_{(i,j) \in E} x_{ij} \leq \sum_{i \leq n} \sum_{j \leq n} x_{ij} \leq \sum_{i \leq n} \sum_j x_{ij} \leq \sum_i \sum_j x_{ij}$; hence $\sum_{ij} x_{ij}$ converges and has sum at most $\sum_i \sum_j x_{ij}$. On the other hand, if $\sum_{ij} x_{ij}$ converges, then $\sum_{i \leq m} \sum_{j \leq n} x_{ij} \leq \sum_{ij} x_{ij}$; letting $n \to \infty$ and then $m \to \infty$ shows that each $\sum_j x_{ij}$ converges and that $\sum_i \sum_j x_{ij} \leq \sum_{ij} x_{ij}$. Therefore, in the nonnegative case, $\sum_{ij} x_{ij}$ converges if and only if the $\sum_j x_{ij}$ all converge and $\sum_i \sum_j x_{ij}$ converges, in which case $\sum_{ij} x_{ij} = \sum_i \sum_j x_{ij}$.

By symmetry, $\sum_{ij} x_{ij} = \sum_j \sum_i x_{ij}$. Thus the order of summation can be reversed in a nonnegative double series: $\sum_i \sum_j x_{ij} = \sum_j \sum_i x_{ij}$.

A28. *The Weierstrass M-test.*

Theorem. *Suppose that* $\lim_n x_{nk} = x_k$ *for each* k *and that* $|x_{nk}| \le M_k$, *where* $\Sigma_k M_k < \infty$. *Then* $\Sigma_k x_k$ *and all the* $\Sigma_k x_{nk}$ *converge, and* $\lim_n \Sigma_k x_{nk} = \Sigma_k x_k$.

PROOF. The series of course converge absolutely, since $\Sigma_k M_k < \infty$. Now $|\Sigma_k x_{nk} - \Sigma_k x_k| \le \Sigma_{k \le k_0} |x_{nk} - x_k| + 2\Sigma_{k > k_0} M_k$. Given ϵ, choose k_0 so that $\Sigma_{k > k_0} M_k < \epsilon/3$, and then choose n_0 so that $n > n_0$ implies $|x_{nk} - x_k| < \epsilon/3k_0$ for $k \le k_0$. Then $n > n_0$ implies $|\Sigma_k x_{nk} - \Sigma_k x_k| < \epsilon$. ∎

A29. *Power series.* The principal fact needed is this: If $f(x) = \Sigma_{k=0}^{\infty} a_k x^k$ converges in the range $|x| < r$, then it is differentiable there and

$$(8) \qquad\qquad f'(x) = \sum_{k=1}^{\infty} ka_k x^{k-1}.$$

For a simple proof, choose r_0 and r_1 so that $|x| < r_0 < r_1 < r$. If $|h| < r_0 - |x|$, so that $|x \pm h| < r_0$, then the mean-value theorem gives (here $0 \le \theta_h \le 1$)

$$(9) \qquad \left| \frac{(x+h)^k - x^k}{h} - kx^{k-1} \right| = \left| k(x + \theta_h h)^{k-1} - kx^{k-1} \right| \le 2kr_0^{k-1}.$$

Since $2kr_0^{k-1}/r_1^k$ goes to 0, it is bounded by some M, and if $M_k = |a_k| \cdot Mr_1^k$, then $\Sigma_k M_k < \infty$ and $|a_k|$ times the left member of (9) is at most M_k for $|h| < r_0 - |x|$. By the M-test [A28] (applied with $h \to 0$ instead of $n \to \infty$),

$$\lim_{h \to 0} \sum_{k=0}^{\infty} a_k \frac{(x+h)^k - x^k}{h} = \sum_{k=0}^{\infty} ka_k x^{k-1}.$$

Hence (8).

Repeated application of (8) gives

$$f^{(j)}(x) = \sum_{k=j}^{\infty} k(k-1) \cdots (k-j+1)a_k x^{k-j}.$$

For $x = 0$, this is $a_j = f^{(j)}(0)/j!$, the formula for the coefficients in a Taylor series. This shows in particular that the values of $f(x)$ for $|x| < r$ determine the coefficients a_k.

A30. *Cesàro averages.* If $x_n \to x$, then $n^{-1}\Sigma_{k=1}^{n} x_k \to x$. To prove this, let M bound $|x_k|$, and given ϵ, choose k_0 so that $|x - x_k| < \epsilon/2$ for $k \ge k_0$. If $n > k_0$ and $n > 4k_0 M/\epsilon$, then

$$\left| x - \frac{1}{n} \sum_{k=1}^{n} x_k \right| \le \frac{1}{n} \sum_{k=1}^{k_0-1} 2M + \frac{1}{n} \sum_{k=k_0}^{n} \frac{\epsilon}{2} < \epsilon.$$

A31. *Dyadic expansions.* Define a mapping T of the unit interval $\Omega = (0, 1]$ into itself by

$$T\omega = \begin{cases} 2\omega & \text{if } 0 < \omega \le \frac{1}{2}, \\ 2\omega - 1 & \text{if } \frac{1}{2} < \omega \le 1. \end{cases}$$

Define a function d_1 on Ω by

$$d_1(\omega) = \begin{cases} 0 & \text{if } 0 < \omega \le \frac{1}{2}, \\ 1 & \text{if } \frac{1}{2} < \omega \le 1, \end{cases}$$

and let $d_i(\omega) = d_1(T^{i-1}\omega)$. Then

(10)
$$\sum_{i=1}^{n} \frac{d_i(\omega)}{2^i} < \omega \le \sum_{i=1}^{n} \frac{d_i(\omega)}{2^i} + \frac{1}{2^n}$$

for all $\omega \in \omega$ and $n \ge 1$. To verify this for $n = 1$, check the cases $\omega \le \frac{1}{2}$ and $\omega > \frac{1}{2}$ separately. Suppose that (10) holds for a particular n and for all ω. Replace ω by $T\omega$ in (10) and use the fact that $d_i(T\omega) = d_{i+1}(\omega)$; separate consideration of the cases $\omega \le \frac{1}{2}$ and $\omega > \frac{1}{2}$ now shows that (10) holds with $n + 1$ in place of n

Thus (10) holds for all n and ω, and it follows that $\omega = \sum_{i=1}^{\infty} d_i(\omega)/2^i$. This gives the dyadic representation of ω. If $d_i(\omega) = 0$ for all $i > n$, then $\omega = \sum_{i=1}^{n} d_i(\omega)/2^i$, which contradicts the left-hand inequality in (10). Thus the expansion does not terminate in 0's.

Convex Functions

A32. A function φ on an open interval I (bounded or unbounded) is *convex* if

(11)
$$\varphi(tx + (1-t)y) \le t\varphi(x) + (1-t)\varphi(y)$$

for $x, y \in I$ and $0 \le t \le 1$. From this it follows by induction that

(12)
$$\varphi\left(\sum_{i=1}^{n} p_i x_i\right) \le \sum_{i=1}^{n} p_i \varphi(x_i)$$

if the x_i lie in I and the p_i are nonnegative and add to 1.

If φ has a continuous, nondecreasing derivative φ' on I, then φ is convex. Indeed, if $a < b < c$, the average of φ' over (a, b) is at most the average over (b, c):

$$\frac{\varphi(b) - \varphi(a)}{b - a} = \frac{1}{b-a}\int_a^b \varphi'(s)\,ds \le \varphi'(b) \le \frac{1}{c-b}\int_b^c \varphi'(s)\,ds$$

$$= \frac{\varphi(c) - \varphi(b)}{c - b}.$$

The inequality between the extreme terms here reduces to

(13)
$$(c - a)\varphi(b) \le (c - b)\varphi(a) + (b - a)\varphi(c),$$

which is (11) with $x = a$, $y = c$, $t = (c - b)/(c - a)$.

A33. Geometrically, (11) means that the point B in Figure (i) lies on or below the chord AC. But then slope $AB \le$ slope AC; algebraically, this is $(\varphi(b) - \varphi(a))/(b - a)$ $\le (\varphi(c) - \varphi(a))/(c - a)$, which is the same as (13). As B moves to A from the right,
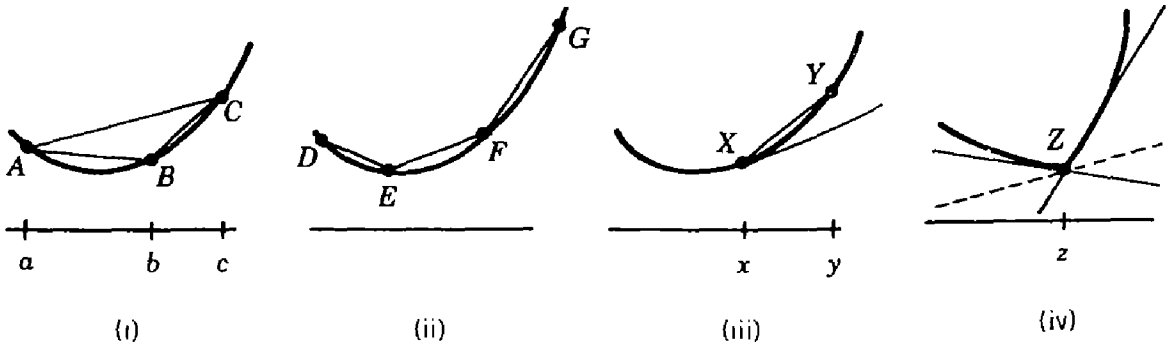
slope AB is thus nonincreasing and hence has a limit. In other words, φ has a right-hand derivative φ^+. Figure (ii) shows that slope $DE \leq$ slope $EF \leq$ slope FG. Let E move to D from the right and let G move to F from the right. The right-hand derivative at D is at most that at F, and φ^+ is nondecreasing. Since the slope of XY in Figure (iii) is at least as great as the right-hand derivative at X, the curve to the right of X lies on or above the line through X with slope $\varphi^+(x)$:

$$(14) \qquad \varphi(y) \geq \varphi(x) + (y-x)\varphi^+(x), \qquad y \geq x.$$

Figure (iii) also makes it clear that φ is right-continuous.

Similarly, φ has a nondecreasing left-hand derivative φ^- and is left-continuous. Since slope $AB \leq$ slope BC in Figure (i), $\varphi^-(b) \leq \varphi^\gamma(b)$. Since clearly $\varphi^+(b) < \infty$ and $-\infty < \varphi^-(b)$, φ^+ and φ^- are finite. Finally, (14) and its right-sided analogue show that the curve lies on or above each line through Z in Figure (iv) having slope between $\varphi^-(z)$ and $\varphi^+(z)$:

$$(15) \qquad \varphi(x) \geq \varphi(z) + m(x-z), \qquad \varphi^-(z) \leq m \leq \varphi^+(z).$$

This is a *support line*.

Some Multivariable Calculus

A34. Suppose that U is an open set in R^k and $T: U \to R^k$ is continuously differentiable; let $D_x = [t_{ij}(x)]$ and $J(x) = \det D_x$ be the Jacobian matrix and determinant, as in Theorem 17.2. Let Q^- be a closed rectangle in U.

Theorem. *If* $|t_{ij}(x') - t_{ij}(x)| \leq \alpha$ *for* $x, x' \in Q^-$ *and all* i, j, *then*

$$(16) \qquad |Tx' - Tx - D_x(x'-x)| \leq k^2\alpha|x'-x|, \qquad x, x' \in Q.$$

Before proceeding to the proof, note that, since the t_{ij} are continuous, α can be taken to go to 0 as Q contracts to the point x. In other words, (16) implies

$$(17) \qquad \lim_{x' \to x} \frac{|Tx' - Tx - D_x(x'-x)|}{|x'-x|} = 0.$$

This shows that D_x acts as a multivariable derivative. Suppose on the other hand that (17) holds at x for an initially unspecified matrix D_x. Take $x'_j = x_j + h$ and $x'_l = x_l$ for $l \neq j$, and let h go to 0. It follows that the entires of D_x *must* be the partial derivatives $t_{ij}(x)$: If (17) holds, then D_x must be the Jacobian matrix.

(v)

(vi)
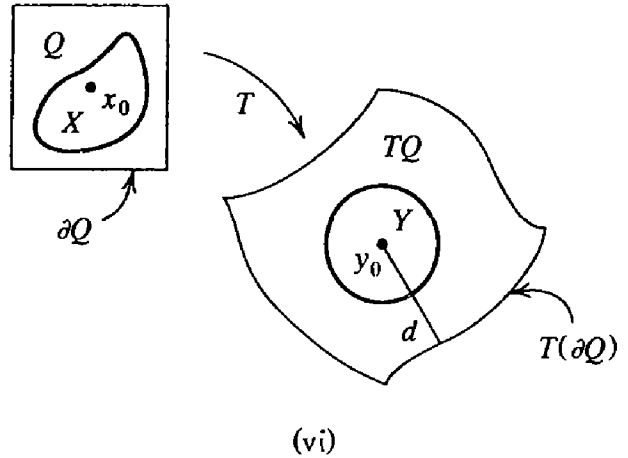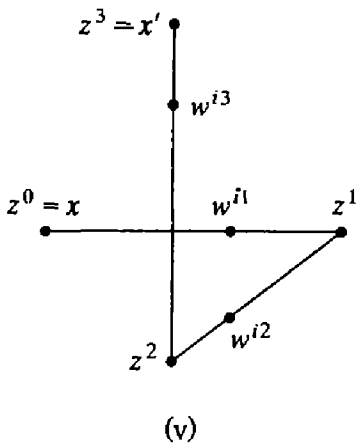
PROOF OF (16).   For $j = 0, 1, \ldots, k$, let z^j agree with x' in the first j places and with x in the last $k - j$ places. Then $z^0 = x$, $z^k = x'$, and $|z^j - z^{j-1}| = |(z^j - z^{j-1})_j| = |x'_j - x_j|$ (Figure (v)). By the mean-value theorem in one dimension, there is a point w^{ij} on the segment from z^{j-1} to z^j such that $t_i(z^j) - t_i(z^{j-1}) = t_{ij}(w^{ij})(z^j - z^{j-1})_j$. Since

$$\left(D_x(z^j - z^{j-1})\right)_i = \sum_l t_{il}(x)(z^j - z^{j-1})_l = t_{ij}(x)(z^j - z^{j-1})_j,$$

it follows that

$$\left|Tx' - Tx - D_x(x' - x)\right|$$

$$\leq \sum_{ij} \left|t_i(z^j) - t_i(z^{j-1}) - \left(D_x(z^j - z^{j-1})\right)_i\right|$$

$$= \sum_{ij} \left|t_{ij}(w^{ij}) - t_{ij}(x)\right| \cdot \left|(z^j - z^{j-1})_j\right|$$

$$\leq \sum_{ij} \alpha |x'_j - x_j| \leq k^2 \alpha |x' - x|. \qquad \blacksquare$$

A35.   *The multivariable inverse-function theorem.* Let x_0 be a point of the open set U.

Theorem.   *If $J(x_0) \neq 0$, then there are open sets X and Y, containing x_0 and $y_0 = Tx_0$, respectively, such that T is a one-to-one map from X onto $Y = TX$; further, $T^{-1}\colon Y \to X$ is continuously differentiable, and the Jacobian matrix of T^{-1} at y is $D_T^{-1}{}_y$.*

This is a local theorem. It is not assumed, as in Theorem 17.2, that T is one-to-one on U and $J(x)$ never vanishes; but under those additional conditions, TU is open and the inverse point mapping is continuously differentiable. To understand the role of the condition $J(x_0) \neq 0$, consider the case where $k = 1$, $x_0 = 0$, and Tx is x^2 or x^3.

PROOF.   Let Q be a rectangle such that $x_0 \in Q^\circ \subset Q^- \subset U$ and $J(x) \neq 0$ for $x \in Q^-$. As (x, u) ranges over the compact set $Q^- \times [u\colon |u| = 1]$, $|D_x u|$ is bounded below by some positive β:

(18) $|D_x u| \geq \beta |u|$ if $x \in Q^-, u \in R^k$.

Making Q smaller will ensure that $|t_{ij}(x) - t_{ij}(x')| \leq \beta/2k^2$ for all x and x' in Q^- and all i, j. Then (16) and (18) give, for $x, x' \in Q^-$,

$$|Tx' - Tx| \geq |D_x(x' - x)| - |Tx' - Tx - D_x(x' - x)|$$

$$\geq |D_x(x' - x)| - \tfrac{1}{2}\beta|x' - x| \geq \tfrac{1}{2}\beta|x' - x|.$$

Thus

(19) $$|x' - x| \leq \frac{2}{\beta}|Tx' - Tx| \qquad \text{for } x, x' \in Q^-.$$

This shows that T is one-to-one on Q^-.

Since x_0 does not lie in the compact set ∂Q, $\inf_{x \in \partial Q}|Tx - Tx_0| = d > 0$. Let Y be the open ball with center $y_0 = Tx_0$ and radius $d/2$ (Figure (vi)). Fix a y in Y. The problem is to show that $y = Tx$ for some x in Q°, which means finding an x such that $\varphi(x) = |y - Tx|^2 = \sum_i(y_i - t_i(x))^2$ vanishes. By compactness, the minimum of φ on Q^- is achieved there. If $x \in \partial Q$ (and $y \in Y$), then $2|y - y_0| < d \leq |Tx - y_0| \leq |Tx - y| + |y - y_0|$, so that $|y - Tx_0| < |y - Tx|$. Therefore, $\varphi(x_0) < \varphi(x)$ for $x \in \partial Q$, and so the minimum occurs in Q° rather than on ∂Q. At the minimizing point, $\partial\varphi/\partial x_j = -\sum_i 2(y_i - t_i(x))t_{ij}(x) = 0$, and since D_x is nonsingular, it follows that $y = Tx$: Each y in Y is the image under T of some point x in Q°. By (19), this x is unique (although it is possible that $y = Tz$ for some z outside Q).

Let $X = Q^\circ \cap T^{-1}Y$. Then X is open and T is a one-to-one map of X onto Y. Now let T^{-1} denote the inverse point transformation on Y. By (19), T^{-1} is continuous.

To prove differentiability, consider in Y a fixed point y and a variable point y' such that $y' \to y$ and $y' \neq y$. Let $x = T^{-1}y$ and $x' = T^{-1}y'$; then x' is a function of y', $x' \to x$, and $x' \neq x$. Define v by $Tx' - Tx = D_x(x' - x) + v$; then v is a function of x' and hence of y', and $|v|/|x' - x| \to 0$ by (17). Apply D_x^{-1}: $D_x^{-1}(Tx' - Tx) = x' - x + D_x^{-1}v$, or $T^{-1}y' - T^{-1}y = D_x^{-1}(y' - y) - D_x^{-1}v$. By (18) and (19),

$$\frac{|T^{-1}y' - T^{-1}y - D_x^{-1}(y' - y)|}{|y' - y|} = \frac{|D_x^{-1}v|}{|x' - x|} \cdot \frac{|x' - x|}{|y' - y|} \leq \frac{|v|/\beta}{|x' - x|} \cdot \frac{2}{\beta}.$$

The right side goes to 0 as $y' \to y$.

By the remark following (17), the components of D_x^{-1} must be the partial derivatives of the inverse mapping. T^{-1} has Jacobian matrix $D_{T^{-1}y}^{-1}$ at y. The components of an inverse matrix vary continuous with the components of the original matrix (think for example of the inverse as specified by the cofactors), and so T^{-1} is even continuously differentiable on Y.   ∎

Continued Fractions

A36. In designing a planetarium, Christian Huygens confronted this problem: Given the ratio x of the periods of two planets, approximate it by the ratio of the periods of two linked gears. If one gear has p teeth and the other has q, then the ratio of their periods is p/q, so that the problem is to approximate the real number x by the rational p/q. Of course x, being empirical, is already rational, but the numerator and denominator may be so large that gears with those numbers of teeth are not practical: in the approximation p/q, both p and q must be of moderate size.

Since the gears play symmetrical roles, there is no more reason to approximate x by $r = p/q$ than there is to approximate $1/x$ by $1/r = q/p$. Suppose, to be definite, that x and r lie to the *left* of 1. Then

$$(20) \qquad |x - r| = xr \left| \frac{1}{x} - \frac{1}{r} \right| \le \left| \frac{1}{x} - \frac{1}{r} \right|,$$

and the inequality is strict unless $x = r$: If $x < 1$, it is better to approximate $1/x$ and then invert the approximation, since that will control *both* errors.

For a numerical illustration, approximate $x = .127$ by rounding it up to $r = .13$; the calculations (to three places) are

$$(21) \qquad \begin{aligned} r - x &= .13 - .127 = .003, \\ \frac{1}{x} - \frac{1}{r} &= 7.874 - 7.692 = .182. \end{aligned}$$

The second error is large. So instead, approximate $1/x = 7.874$ by rounding it down to $1/r' = 7.87$. Since $1/7.87 = .1271$ (to four places), the calculations are

$$(22) \qquad \begin{aligned} \frac{1}{x} - \frac{1}{r'} &= 7.874 - 7.87 = .004, \\ r' - x &= .1271 - .127 = .0001. \end{aligned}$$

This time, both errors are small, and the error .0001 in the new approximation to x is smaller than the corresponding .003 in (21). It is because x lies to the *left* of 1 that inversion improves the accuracy; see (20).

If this inversion method decreases the error, why not do another inversion in the middle, in finding a rational approximation to $1/x$? It makes no sense to invert $1/x$ itself, since it lies to the right of 1 (and inversion merely leads back to x anyway). But to approximate $1/x = 7.874$ is to approximate the fractional part .874, and here a second inversion will help, for the same reason the first one does. This suggests Huygens's iterative procedure.

In modern notation, the scheme is this. For x (rational or irrational) in $(0, 1)$, let $Tx = \{1/x\}$ and $a_1(x) = \lfloor 1/x \rfloor$ be the fractional and integral parts of $1/x$; and set $T0 = 0$. This defines a mapping of $[0, 1)$ onto itself:

$$(23) \qquad Tx = \begin{cases} \left\{ \dfrac{1}{x} \right\} = \dfrac{1}{x} - \left\lfloor \dfrac{1}{x} \right\rfloor = \dfrac{1}{x} - a_1(x) & \text{if } 0 < x < 1, \\ 0 & \text{if } x = 0. \end{cases}$$

Then

$$(24) \qquad x = \frac{1}{a_1(x) + Tx} \qquad \text{if } 0 < x < 1.$$

What (20) says is that replacing Tx on the right in (24) by a good rational approximation to it gives an even better rational approximation to x itself.

To carry this further requires a convenient notation. For positive variables z_i define the *continued fractions*

$$\underline{1|z_1} = \frac{1}{z_1}, \qquad \underline{1|z_1} + \underline{1|z_2} = \cfrac{1}{z_1 + \cfrac{1}{z_2}},$$

$$\underline{1|z_1} + \underline{1|z_2} + \underline{1|z_3} = \cfrac{1}{z_1 + \cfrac{1}{z_2 + \cfrac{1}{z_3}}},$$

and so on. It is "typographically" clear that

(25)
$$\underline{1|z_1} + \cdots + \underline{1|z_n} = 1 \Big/ \Big[z_1 + \big(\underline{1|z_2} + \cdots + \underline{1|z_n} \big) \Big]$$

and

(26)
$$\underline{1|z_1} + \cdots + \underline{1|z_n} = \underline{1|z_1} + \cdots + \underline{1|z_{n-1} + 1/z_n}.$$

For a formal theory, use (25) as a recursive definition and then prove (26) by induction (or vice versa). An infinite continued fraction is defined by

$$\underline{1|z_1} + \underline{1|z_2} + \cdots = \lim_n \underline{1|z_1} + \cdots + \underline{1|z_n},$$

provided the limit exists. A continued fraction is *simple* if the z_i are positive integers. If $T^{n-1}x > 0$, let $a_n(x) = a_1(T^{n-1}x)$; the $a_n(x)$ are the *partial quotients* of x. If x and Tx are both positive, then (24) applies to each of them:

$$x = \underline{1|a_1(x) + Tx} = \underline{1|a_1(x)} + \underline{1|a_2(x) + T^2 x}.$$

If none of the iterates $x, Tx, \ldots, T^{n-1}x$ vanishes, then it follows by induction (use (26)) that

(27)
$$x = \underline{1|a_1(x)} + \cdots + \underline{1|a_{n-1}(x)} + \underline{1|a_n(x) + T^n x}.$$

This is an extension of (24), and the idea following (24) extends as well: a good rational approximation to $T^n x$ in (27) gives a still better rational approximation to x. Even if $T^n x$ is approximated very crudely by 0, there results a sharp approximation

(28)
$$x \approx \underline{1|a_1(x)} + \cdots + \underline{1|a_n(x)}$$

to x itself. The right side here is the nth *convergent* to x, and it goes very rapidly to x; see Section 24.

By the definition (23), x and Tx are both rational or both irrational. For an irrational x, therefore, $T^n x$ remains forever among the irrationals, and (27) holds for all n. If x is rational, on the other hand, $T^n x$ remains forever among the rationals, and in fact, as the following argument shows, $T^n x$ eventually hits 0 and stays there.

Suppose that x is a rational in $(0,1)$: $x = d_1/d_0$, where $0 < d_1 < d_0$. If $Tx > 0$, then $Tx = \{d_0/d_1\} = d_2/d_1$, where $0 < d_2 < d_1$ because $0 < Tx < 1$. (If d_1/d_0 is irreducible, so is d_2/d_1.) If $T^2x > 0$, the argument can be repeated:

$$(29) \qquad x = \frac{d_1}{d_0}, \quad Tx = \frac{d_2}{d_1}, \quad T^2x = \frac{d_3}{d_2}, \qquad d_0 > d_1 > d_2 > d_3 > 0.$$

And so on. Since the d_n decrease as long as the $T^n x$ remain positive, $T^n x$ must vanish for some n, and then $T^m x = 0$ for $m \geq n$. If n_x is the smallest integer for which $T^n x = 0$ $(n_x \geq 1$ if $x > 0)$, then by (27),

$$(30) \qquad x = 1\!\!\left\lfloor a_1(x) + \cdots + 1\!\!\left\lfloor a_{n_x}(x). \right.\right.$$

Thus each positive rational has a representation as a finite simple continued fraction. If $0 < x < 1$ and $Tx = 0$, then $1 > x = 1/a_1(x)$, so that $a_1(x) \geq 2$. Applied to $T^{n_x-1}x$, this shows that the $a_{n_x}(x)$ in (30) must be at least 2.

Section 24 requires a uniqueness result. Suppose that

$$(31) \qquad x = 1\!\!\left\lfloor a_1 + \cdots + 1\!\!\left\lfloor a_{n-1} + 1\!\!\left\lfloor a_n + t, \right.\right.\right.$$

where the a_i are positive integers and

$$(32) \qquad 0 < x < 1, \qquad 0 \leq t < 1, \qquad a_n + t > 1.$$

The last condition rules out $a_n = 1$ and $t = 0$ (which in the case $n = 1$ is also ruled out by $x < 1$). It follows from (31) and (32) that

$$(33) \qquad a_1(x) = a_1, \ldots, a_n(x) = a_n, \qquad T^n x = t.$$

The case $n = 1$ being easy, suppose the implication holds for $n - 1$, where $n \geq 2$. Since $0 < 1/(a_n + t) < 1$, the induction hypothesis (use (26)) gives $a_k(x) = a_k$ for $k < n$ and $T^{n-1}x = 1/(a_n + t)$. Now apply the case $n = 1$ to $T^{n-1}x$. (If $a_n = 1$ and $t = 0$, then $a_k(x) = a_k$ for $k \leq n - 2$, $a_{n-1}(x) = a_{n-1} + 1$, and $T^{n-1}x = 0$.)

Consider now the infinite case. Assume that

$$(34) \qquad x = 1\!\!\left\lfloor a_1 + 1\!\!\left\lfloor a_2 + \cdots , \right.\right.$$

converges, where the a_n are positive integers. Then

$$(35) \qquad a_n(x) = a_n, \qquad T^n x = 1\!\!\left\lfloor a_{n+1} + 1\!\!\left\lfloor a_{n+2} + \cdots , \qquad n \geq 1.\right.\right.$$

To prove this, let $n \to \infty$ in (25): the continued fraction $t = 1\!\!\left\lfloor a_2 + 1\!\!\left\lfloor a_3 + \cdots \right.\right.$ converges and $x = 1/(a_1 + t)$. It follows by induction (use (26)) that

$$(36) \qquad 1\!\!\left\lfloor a_1 > 1\!\!\left\lfloor a_1 + \cdots + 1\!\!\left\lfloor a_n \geq 1\!\!\left\lfloor a_1 + 1\!\!\left\lfloor a_2, \qquad n \geq 2.\right.\right.\right.\right.\right.$$

Hence $0 < x < 1$, and the same must be true of t. Therefore, a_1 and t are the integer and fractional parts of $1/x$, which proves (35) for $n = 1$. Apply the same argument to

Tx, and continue. The x defined by (34) is irrational: otherwise, $T^n x = 0$ for some n, which contradicts (35) and (36).

Thus the value of an infinite simple continued fraction uniquely determines the partial quotients. The same is almost true of finite simple continued fractions. Since (31) and (32) imply (33), it follows that if x is given by (30), then any continued fraction of n_x terms that represents x must indeed match (30) term for term. But, for example, $1\lceil 3 + 1 \rceil 5 = 1 \lceil 3 + 1 \lceil 4 + 1 \lceil 1$. This is always possible: replace $a_{n_x}(x)$ in (30) (where $a_{n_x}(x) \geq 2$) by $a_{n_x}(x) - 1 + 1\lceil 1$. Apart from this ambiguity, the representation is unique—and the representation (30) that results from repeated application of T to a rational x never ends with a partial quotient of 1.[†]

[†]See Rockett & Szusz for more on continued fractions.

Notes on the Problems

These notes consist of hints, solutions, and references to the literature. As a rule a solution is complete in proportion to the frequency with which it is needed for the solution of subsequent problems

Section 1

1.1. (a) Each point of the discrete space lies in one of the four sets $A_1 \cap A_2$, $A_1^c \cap A_2$, $A_1 \cap A_2^c$, $A_1^c \cap A_2^c$ and hence would have probability at most 2^{-2}; continue.

(b) If, for each i, B_i is A_i or A_i^c, then $B_1 \cap \cdots \cap B_n$ has probability at most $\prod_{i=1}^n (1 - \alpha_i) \leq \exp[-\sum_{i=1}^n \alpha_i]$

1.3. (b) Suppose A is trifling and let A^- be its closure. Given ϵ choose intervals $(a_k, b_k]$, $k = 1, \ldots, n$, such that $A \subset \bigcup_{k=1}^n (a_k, b_k]$ and $\sum_{k=1}^n (b_k - a_k) < \epsilon/2$. If $x_k = a_k - \epsilon/2n$, then $A^- \subset \bigcup_{k=1}^n (x_k, b_k]$ and $\sum_{k=1}^n (b_k - x_k) < \epsilon$.

For the other parts of the problem, consider the set of rationals in $(0, 1)$.

1.4. (a) Cover $A_r(i)$ by $(r - 1)^n$ intervals of length r^{-n}.

(c) Go to the base r^k. Identify the digits in the base r with the keys of the typewriter. The monkey is certain eventually to reproduce the eleventh edition of the *Britannica* and even, unhappily, the fifteenth.

1.5. (a) The set $A_3(1)$ is itself uncountable, since a point in it is specified by a sequence of 0's and 2's (excluding the countably many that end in 0's).

(b) For sequences $u_1, \ldots, u_n$ of 0's, 1's, and 2's, let $M_{u_1 \ldots u_n}$ consist of the points in $(0, 1]$ whose nonterminating base-3 expansions start out with those digits. Then $A_3(1) = (0, 1] - \bigcup M_{u_1 \ldots u_n}$, where the union extends over $n \geq 1$ and the sequences $u_1, \ldots, u_n$ containing at least one 1. The set described in part (b) is $[0, 1] - \bigcup M_{u_1 \ldots u_n}^o$, where the union is as before, and this is the closure of $A_3(1)$.

From this representation of C, it is not hard to deduce that it can be defined as the set of points in $[0, 1]$ that can be written in base 3 without any 1's if terminating expansions are also allowed. For example, C contains $\frac{2}{3} = .1222 \cdots = .2000\ldots$ because it is possible to avoid 1 in the expansion.

(c) Given ϵ and an ω in C, choose ω' in $A_3(1)$ within $\epsilon/2$ of ω; now define ω'' by changing from 2 to 0 some digit of ω' far enough out that ω'' differs from ω' by at most $\epsilon/2$.

1.7. The interchange of limit and integral is justified because the series $\sum_k r_k(\omega)2^{-k}$ converges uniformly in ω (integration to the limit is studied systematically in Section 16). There is a direct derivation of (1.40): let $n \to \infty$ in $\sin t = 2^n \sin 2^{-n} t \cdot \prod_{k=1}^n \cos 2^{-k} t$, which follows by induction from the half-angle formula.

1.10. (a) Given m and a subinterval $(a, b]$ of $(0,1]$, choose a dyadic interval I in $(a, b]$, and then choose in I a dyadic interval J of order $n > m$ such that $|n^{-1}s_n(\omega)| > \frac{1}{2}$ for $\omega \in J$. This is possible because to specify J is to specify the first n dyadic digits of the points in J, choose the first digits in such a way that $J \subset I$ and take the following ones to be 1, with n so large that $n^{-1}s_n(\omega)$ is near 1 for $\omega \in J$.

(b) A countable union of sets of the first category is also of the first category; $(0, 1] = N \cup N^c$ would be of the first category if N^c were. For Baire's theorem, see ROYDEN, p 139.

1.11. (a) If $x = p_0/q_0 \neq p/q$, then

$$\left| x - \frac{p}{q} \right| = \frac{|p_0 q - q_0 p|}{q_0 q} \geq \frac{1}{q_0 q}.$$

(c) The rational $\sum_{k=1}^n 1/2^{\alpha(k)}$ has denominator $2^{\alpha(n)}$ and approximates x to within $2/2^{\alpha(n+1)}$.

Section 2

2.3. (b) Let Ω consist of four points, and let $\mathscr{F}$ consist of the empty set, Ω itself, and all six of the two-point sets.

2.4. (b) For example, take Ω to consist of the integers, and let $\mathscr{F}_n$ be the σ-field generated by the singletons $\{k\}$ with $k \leq n$. As a matter of fact, any example in which $\mathscr{F}_n$ is a proper subclass of $\mathscr{F}_{n+1}$ for all n will do, because it can be shown that in this case $\bigcup_n \mathscr{F}_n$ necessarily fails to be a σ-field; see A. Broughton and B. W. Huff: A comment on unions of sigma-fields, *Amer. Math. Monthly*, **84** (1977), 553–554.

2.5. (b) The class in question is certainly contained in $f(\mathscr{A})$ and is easily seen to be closer under the formation of finite intersections. But $(\bigcup_{i=1}^m \bigcap_{j=1}^n A_{ij})^c = \bigcap_{i=1}^m \bigcup_{j=1}^n A_{ij}^c$, and $\bigcup_{j=1}^n A_{ij}^c = \bigcup_{j=1}^n [A_{ij}^c \cap \bigcap_{k=1}^{j-1} A_{ik}]$ has the required form.

2.8. If $\mathscr{H}$ is the smallest class over $\mathscr{A}$ closed under the formation of countable unions and intersections, clearly $\mathscr{H} \subset \sigma(\mathscr{A})$. To prove the reverse inclusion, first show that the class of A such that $A^c \in \mathscr{H}$ is closed under the formation of countable unions and intersections and contains $\mathscr{A}$ and hence contains $\mathscr{H}$.

2.9. Note that $\bigcup_n B_n \in \sigma(\bigcup_n \mathscr{A}_{B_n})$.

2.10. (a) Show that the class of A for which $I_A(\omega) = I_A(\omega')$ is a σ-field. See Example 4.8.

2.11. (b) Suppose that $\mathscr{F}$ is the σ-field of the countable and the cocountable sets in Ω. Suppose that $\mathscr{F}$ is countably generated and Ω is uncountable. Show that $\mathscr{F}$

is generated by a countable class of singletons; if Ω_0 is the union of these, then $\mathscr{F}$ must consist of the sets B and $B \cup \Omega_0^c$ with $B \subset \Omega_0$, and these do not include the singletons in Ω_0^c, which is uncountable because Ω is.

(c) Let $\mathscr{F}_2$ consist of the Borel sets in $\Omega = (0, 1]$, and let $\mathscr{F}_1$ consist of the countable and the cocountable sets there.

2.12. Suppose that $A_1, A_2, \ldots$ is an infinite sequence of distinct sets in a σ-field $\mathscr{F}$, and let $\mathscr{G}$ consist of the nonempty sets of the form $\bigcap_{n=1}^{\infty} B_n$, where $B_n = A_n$ or $B_n = A_n^c$, $n = 1, 2, \ldots$. Each A_n is the union of the $\mathscr{G}$-sets it contains, and since the A_n are distinct, $\mathscr{G}$ must be infinite. But there are uncountably many distinct countable unions of $\mathscr{G}$-sets, and they all lie in $\mathscr{F}$.

2.18. For this and the subsequent problems on applications of probability theory to arithmetic, the only number theory required is the fundamental theorem of arithmetic and its immediate consequences. The other problems on stochastic arithmetic are 4.15, 4.16, 5.19, 5.20, 6.16, 18.17, 25.15, 30.9, 30.10, 30.11, and 30.12. See also Theorem 30.3.

(b) Let A consist of the even integers, let $C_k = [m: v_k < m \le v_{k+1}]$, and let B consist of the even integers in $C_1 \cup C_3 \cup \cdots$ together with the odd integers in $C_2 \cup C_4 \cup \cdots$; take v_k to increase very rapidly with k and consider $A \cap B$.

(c) If c is the least common multiple of a and b, then $M_a \cap M_b = M_c$. From $M_a \in \mathscr{D}$ conclude in succession that $M_a \cap M_b \in \mathscr{D}$, $M_{a_1} \cap \cdots \cap M_{a_j} \cap M_{b_1}^c \cap \cdots \cap M_{b_k}^c \in \mathscr{D}$, $f(\mathscr{A}') \subset \mathscr{D}$. By the same sequence of steps, show how D on $\mathscr{A}$ determines D on $f(\mathscr{A})$.

(d) If $B_l = M_a - \bigcup_{p \le l} M_{ap}$, then $a \in B_l$ and (the inclusion-exclusion formula requires only finite additivity)

$$D(B_l) = \frac{1}{a} - \sum_{p \le l} \frac{1}{ap} + \sum_{p < q \le l} \frac{1}{apq} - \cdots$$

$$= \frac{1}{a} \prod_{p \le l} \left(1 - \frac{1}{p}\right) \le \frac{1}{a} \exp\left(-\sum_{p \le l} \frac{1}{p}\right) \to 0.$$

Choose l_a so that, if $C_a = B_{l_a}$, then $D(C_a) < 2^{-a-1}$. If D were a probability measure on $f(\mathscr{A})$, $D(\Omega) \le \frac{1}{2}$ would follow. See Problem 4.16 for a different approach.

2.19. (a) Apply the intermediate-value theorem to the function $f(x) = \lambda(A \cap (0, x])$. Note that this even proves part (c) for λ (under the assumption that λ exists).

(b) If $0 < P(B) < P(A)$, then either $0 < P(B) \le \frac{1}{2}P(A)$ or $0 < P(B - A) \le \frac{1}{2}P(A)$. Continue.

(c) If $P(\bigcup_k H_k) < x$, choose C so that $C \subset A - \bigcup_k H_k$ and $0 < P(C) < x - P(\bigcup_k H_k)$. If $n^{-1} < P(C)$, then $P(\bigcup_{k<n} H_k) + h_n < P(\bigcup_{k<n} H_k) + P(H_n) + P(C) \le P(\bigcup_{k<n} H_k) + h_n$.

2.21. (c) If $\mathscr{I}_{n-1}$ were a σ-field, $\mathscr{B} \subset \mathscr{I}_{n-1}$ would follow.

2.22. Use the fact that, if $\alpha_1, \alpha_2, \ldots$ is a sequence of ordinals satisfying $\alpha_n < \Omega$, then there exists an ordinal α such that $\alpha < \Omega$ and $\alpha_n < \alpha$ for all n.

2.23. Suppose that $B_j \in \bigcup_{\beta < \alpha} \mathcal{J}_\beta$, $j = 1, 2, \ldots$. Choose odd integers n_j in such a way that $B_j \in \mathcal{J}_{\beta_\alpha(n_j)}$ and the n_j are all distinct; choose $\mathcal{J}_0$-sets such that

$$\Phi_{\beta_\alpha(n_j)}\left(C_{m_{n_j 1}}, C_{m_{n_j 2}}, \ldots\right) = B_j;$$

for n not of the form n_j, choose $\mathcal{J}_0$-sets for which $\Phi_{\beta_\alpha(n)}(C_{m_{n1}}, C_{m_{n2}}, \ldots)$ is $\varnothing$ or $(0, 1]$ as n is odd or even. Then $\bigcup_{j=1}^\infty B_j = \Phi_\alpha(C_1, C_2, \ldots)$. Similarly, $B^c = \Phi_\alpha(C_1, C_2, \ldots)$ for $\mathcal{J}_0$-sets C_n if $B \in \bigcup_{\beta < \alpha} \mathcal{J}_\beta$. The rest of the proof is essentially the same as before.

Section 3

3.1. (a) The finite additivity of P is used in the proof that $\mathcal{F}_0 \subset \mathcal{M}$ and again (via monotonicity; see (2.5)) in the proof of (3.7). The countable additivity of P is used (via countable subadditivity; see Theorem 2.1) in the proof of (3.7).

(b) For a specific example consider $\bigcup_n (2^{-1} + n^{-1}, 1]$ in connection with Problem 2.15. But an example is provided by *every* P that is finitely but not countably additive: If P is finitely additive on a field $\mathcal{F}_0$ and A_n are disjoint $\mathcal{F}_0$-sets whose union A also lies in $\mathcal{F}_0$, then monotonicity (which requires finite additivity only) gives $\sum_{k \le n} P(A_k) = P(\bigcup_{k \le n} A_k) \le P(A)$ and hence $\sum_k P(A_k) \le P(A)$. Countable subadditivity will ensure that there is equality here.

(c) The proof of (3.7) involves the countable subadditivity of P on $\mathcal{F}_0$, which is only assumed to be a field (that being the whole point of the theorem).

3.2. (a) Given ϵ, choose $\mathcal{F}_0$-sets A_n such that $A \subset \bigcup_n A_n$ and $\sum P(A_n) < P^*(A) + \epsilon$; if $B = \bigcup_n A_n$, then $A \subset B$, $B \in \mathcal{F}$, and $P(B) < P^*(A) + \epsilon$; hence the right side of (3.9) is at most $P^*(A)$. On the other hand, $A \subset B$ and $B \in \mathcal{F}$ imply $P^*(A) \le P^*(B) = P(B)$. Hence (3.9). If $A \subset B_k$, $B_k \in \mathcal{F}$, $P(B_k) < P^*(A) + k^{-1}$, and $B = \bigcap_k B_k$, then $A \subset B$, $B \in \mathcal{F}$, and $P^*(A) = P(B)$. For (3.10), argue by complementation.

(b) Suppose that $P_*(A) = P^*(A)$ and chose $\mathcal{F}$-sets A_1 and A_2 in such a way that $A_1 \subset A \subset A_2$ and $P(A_1) = P(A_2)$. Given E, choose an $\mathcal{F}$-set B in such a way that $E \subset B$ and $P^*(E) = P(B)$. Then $P^*(A \cap E) + P^*(A^c \cap E) \le P(A_2 \cap B) + P(A_1^c \cap B)$. Now use (2.7) to bound the last sum by $P(B) + P(A_2 - A_1) = P^*(E)$.

3.3. First note the general fact that P^* agrees with P on $\mathcal{F}_0$ if and *only if* P is countably additive there, a condition not satisfied in parts (b) and (e). Using Problem 3.2 simplifies the analysis of P^* and $\mathcal{M}(P^*)$ in the other parts.

Note in parts (b) and (e) that, if P^* and P_* are defined by (3.1) and (3.2), then, since $P^*(A) = 0$ for all A, (3.4) holds for all A and (3.3) holds for no A. Countable additivity thus plays an essential role in Problem 3.2.

3.6 (c) Split E^c by A: $P_0(E) = 1 - P^0(E^c) = 1 - P^0(A \cap E^c) - P^0(A^c \cap E^c) = 1 - P^0(A \cap E^c) - P(A^c) = P(A) - P^0(A - E)$.

3.7. (b) Apply (3.13): For $A \in \mathcal{F}_0$, $Q(A) = P^0(H \cap A) + P_0(H^c \cap A) = P^0(H \cap A) + P(A) - P^0(A - (H^c \cap A)) = P(A)$.

(c) If A_1 and A_2 are disjoint $\mathcal{F}_0$-sets, then by (3.12),

$$P^0(H \cap (A_1 \cup A_2)) = P^0(H \cap A_1) + P^0(H \cap A_2).$$

Apply (3.13) to the three terms in this equation, successively using $A_1 \cup A_2$, A_1, and A_2 for A:

$$P_0(H^c \cap (A_1 \cup A_2)) = P_0(H^c \cap A_1) + P_0(H^c \cap A_2).$$

But for these two equations to hold it is enough that $H \cap A_1 \cap A_2 = \varnothing$ in the first case and $H^c \cap A_1 \cap A_2 = \varnothing$ in the second (replacing A_1 by $A_1 \cap A_2^c$ changes nothing).

3.8. By using Banach limits (BANACH, p. 34) one can similarly prove that density D on the class $\mathscr{D}$ (Problem 2.18) extends to a finitely additive probability on the class of all subsets of $\Omega = \{1, 2, \dots\}$.

3.14. The argument is based on cardinality. Since the Cantor set C has Lebesgue measure 0, 2^C is contained in the class $\mathscr{L}$ of Lebesgue sets in $(0, 1]$. But C is uncountable: card $\mathscr{B}$ = card$(0, 1]$ < card 2^C ≤ card $\mathscr{L}$.

3.18. (a) Since the $A \oplus r$ are disjoint Borel sets, $\sum_r \lambda(A \oplus r) \le 1$, and so the common value $\lambda(A)$ of the $\lambda(A \oplus r)$ must be 0. Similarly, if A is a Borel set contained in some $H \oplus r$, then $\lambda(A) = 0$.

(b) If the $E \cap (H \oplus r)$ are all Borel sets, they all have Lebesgue measure 0, and so E is a Borel set of Lebesgue measure 0.

3.19. (b) Given $A_1, B_1, \dots, A_{n-1}, B_{n-1}$, note that their union C_n is nowhere dense, so that I_n contains an interval J_n disjoint from C_n. Choose in J_n disjoint, nowhere dense sets A_n and B_n of positive measure.

(c) Note that A and B_n are disjoint and that $A_n \cup B_n \subset G$.

3.20. (a) If I_n are disjoint open intervals with union G, then $b^{-1}\lambda(A) > \sum_n \lambda(I_n) \ge \sum_n b^{-1}\lambda(A \cap I_n) \ge b^{-1}\lambda(A)$.

Section 4

4.1. Let r be the quantity on the right in (4.30), assumed finite. Suppose that $x < r$; then $x < \bigvee_{k=n}^{\infty} x_k$ for $n \ge 1$ and hence $x < x_k$ for some $k \ge n$: $x < x_n$ i.o. Suppose that $x < x_n$ i.o.; then $x < \bigvee_{k=n}^{\infty} x_k$ for $n \ge 1$: $x \le r$. It follows that $r = \sup[x: x < x_n$ i.o.$]$, which is easily seen to be the supremum of the limit points of the sequence. The argument for (4.31) is similar.

4.10. The class $\mathscr{F}$ is the σ-field generated by $\mathscr{G} \cup \{H\}$ (Problem 2.7(a)). If $(H \cap G_1) \cup (H^c \cap G_2) = (H \cap G_1') \cup (H^c \cap G_2')$, then $G_1 \triangle G_1' \subset H^c$ and $G_2 \triangle G_2' \subset H$; consistency now follows because $\lambda_*(H) = \lambda_*(H^c) = 0$. If $A_n = (H \cap G_1^{(n)}) \cup (H^c \cap G_2^{(n)})$ are disjoint, then $G_1^{(m)} \cap G_1^{(n)} \subset H^c$ and $G_2^{(m)} \cap G_2^{(n)} \subset H$ for $m \ne n$, and therefore (see Problem 2.17) $P(\bigcup_n A_n) = \frac{1}{2}\lambda(\bigcup_n G_1^{(n)}) + \frac{1}{2}\lambda(\bigcup_n G_2^{(n)}) = \sum_n (\frac{1}{2}\lambda(G_1^{(n)}) + \frac{1}{2}\lambda(G_2^{(n)})) = \sum_n P(A_n)$. The intervals with rational endpoints generate $\mathscr{G}$.

4.14. Show as in Problem 1.1(b) that the maximum of $P(B_1 \cap \cdots \cap B_n)$, where B_i is A_i or A_i^c, goes to 0. Let $A_x = [\omega: \sum_n I_{A_n}(\omega)2^{-n} \le x]$, show that $P(A \cap A_x)$ is continuous in x, and proceed as in Problem 2.19(a).

4.15. Calculate $D(F_l)$ by (2.36) and the inclusion–exclusion formula, and estimate $P_n(F_l - F)$ by subadditivity; now use $0 \le P_n(F_l) - P_n(F) = P_n(F_l - F)$. For the calculation of the infinite product, see HARDY & WRIGHT, p. 246.

Section 5

5.5. (a) If $m = 0$, $\alpha \ge 0$, and $x > 0$, then $P[X \ge \alpha] \le P[(X + x)^2 \ge (\alpha + x)^2] \le E[(X + x)^2]/(\alpha + x)^2 = (\sigma^2 + x^2)/(\alpha + x)^2$; minimize over x.

5.8. (b) It is enough to prove that $\varphi(t) = f(t(x', y') + (1 - t)(x, y))$ is convex in t $(0 \le t \le 1)$ for (x, y) and (x', y') in C. If $\alpha = x' - x$ and $\beta = y' - y$, then (if $f_{11} > 0$)

$$\varphi'' = f_{11}\alpha^2 + 2f_{12}\alpha\beta + f_{22}\beta^2$$

$$= \frac{1}{f_{11}}(f_{11}\alpha + f_{12}\beta)^2 + \frac{1}{f_{11}}(f_{11}f_{22} - f_{12}^2)\beta^2 \ge 0.$$

Examples like $f(x, y) = y^2 - 2xy$ show that convexity in each variable separately does not imply convexity.

5.9. Check (5.39) for $f(x, y) = -x^{1/p}y^{1/q}$.

5.10. Check (5.39) for $f(x, y) = -(x^{1/p} + y^{1/p})^p$.

5.19. For (5.43) use (2.36) and the fundamental theorem of arithmetic: since the p_i are distinct, the $p_i^{k_i}$ individually divide m if and only if their product does. For (5.44) use inclusion–exclusion. For (5.47), use (5.29) (see Problem 5.12)).

5.20. (a) By (5.47), $E_n[\alpha_p] \le \sum_{k=1}^{\infty} p^{-k} \le 2/p$. And, of course, $n^{-1}\log n! = E_n[\log] = \sum_p E_n[\alpha_p]\log p$.

(b) Use (5.48) and the fact that $E_n[\alpha_p - \delta_p] \le \sum_{k=2}^{\infty} p^{-k}$.

(c) By (5.49),

$$\sum_{n < p \le 2n} \log p = \sum_{n < p \le 2n} \left(\left\lfloor \frac{2n}{p} \right\rfloor - 2\left\lfloor \frac{n}{p} \right\rfloor \right) \log p$$

$$\le 2n(E_{2n}[\log^*] - E_n[\log^*]) = O(n).$$

Deduce (5.50) by splitting the range of summation by successive powers of 2.

(d) If K bounds the $O(1)$ terms in (5.51), then

$$\sum_{p \le x} \log p \ge \theta x \sum_{\theta x < p \le x} p^{-1}\log p \ge \theta x(\log \theta^{-1} - 2K).$$

(e) For (5.53) use

$$\sum_{p \le x} \frac{\log p}{\log x} \le \pi(x) \le \sum_{p \le x^{1/2}} 1 + \sum_{x^{1/2} < p \le x} \frac{\log p}{\log x^{1/2}}$$

$$\le x^{1/2} + \frac{2}{\log x} \sum_{p \le x} \log p.$$

By (5.53), $\pi(x) \geq x^{1/2}$ for large x, and hence $\log \pi(x) \asymp \log x$ and $\pi(x) \asymp x/\log \pi(x)$. Apply this with $x = p_r$ and note that $\pi(p_r) = r$.

Section 6

6.3. Since for given values of $X_{n1}(\omega), \ldots, X_{n,k-1}(\omega)$ there are for $X_n(\omega)$ the k possible values $0, 1, \ldots, k-1$, the number of values of $(X_{n1}(\omega), \ldots, X_{nn}(\omega))$ is $n!$. Therefore, the map $\omega \to (X_{n1}(\omega), \ldots, X_{nn}(\omega))$ is one-to-one, and the $X_{nk}(\omega)$ determine ω. It follows that if $0 \leq x_i < i$ for $1 \leq i \leq k$, then the number of permutations ω satisfying $X_{ni}(\omega) = x_i$, $1 \leq i \leq k$, is just $(k+1)(k+2) \cdots n$, so that $P[X_{ni} = x_i, 1 \leq i \leq k] = 1/k!$. It now follows by induction on k that $X_{n1}, \ldots, X_{nk}$ are independent and $P[X_{nk} = x] = k^{-1}$ $(0 \leq x < k)$.

Now calculate

$$E[X_{nk}] = \frac{k-1}{2},$$

$$E[S_n] = \frac{0 + 1 + \cdots + (n-1)}{2} = \frac{n(n-1)}{4} \sim \frac{n^2}{4},$$

$$\mathrm{Var}[X_{nk}] = \frac{0^2 + 1^2 + \cdots + (k-1)^2}{k} - \left(\frac{k-1}{2}\right)^2 = \frac{k^2 - 1}{12},$$

$$\mathrm{Var}[S_n] = \frac{1}{12} \sum_{k=1}^{n} (k^2 - 1) = \frac{2n^3 + 3n^2 - 5n}{72} \sim \frac{n^3}{36}.$$

Apply Chebyshev's inequality.

6.7. (a) If $k^2 \leq n < (k+1)^2$, let $a_n = k^2$; if M bounds the $|x_n|$, then

$$\left| \frac{1}{n} s_n - \frac{1}{a_n} s_{a_n} \right| \leq \left| \frac{1}{n} - \frac{1}{a_n} \right| \cdot nM + \frac{1}{a_n}(n - a_n)M = 2M \frac{n - a_n}{a_n} \to 0.$$

6.16. From (5.53) and (5.54) it follows that $a_n = \sum_p n^{-1} \lfloor n/p \rfloor \to \infty$. The left side of (6.8) is

$$\frac{1}{n}\left\lfloor \frac{n}{pq} \right\rfloor - \frac{1}{n}\left\lfloor \frac{n}{p} \right\rfloor \frac{1}{n}\left\lfloor \frac{n}{q} \right\rfloor \leq \frac{1}{pq} - \left(\frac{1}{p} - \frac{1}{n}\right)\left(\frac{1}{q} - \frac{1}{n}\right) \leq \frac{1}{np} + \frac{1}{nq}.$$

Section 7

7.3. If one grants that there are only countably many effective rules, the result is an immediate consequence of the mathematics of this and the preceding sections: C is a countable intersection of $\mathscr{F}$-sets of measure 1. The argument proves in particular the nontrivial fact that collectives exist.

7.7. If $n \leq \tau$, then $W_n = W_{n-1} - X_{n-1} = W_1 - S_{n-1}$, and τ is the smallest n for which $S_{n-1} = W_1$. Use (7.8) for the question of whether the game terminates. Now

$$F_\tau = F_0 + \sum_{k=1}^{\tau-1} (W_1 - S_{k-1}) X_k = F_0 + W_1 S_{\tau-1} - \tfrac{1}{2}\left(S_{\tau-1}^2 - (\tau - 1)\right).$$

7.8. Let $x_1, \ldots, x_i$ be the initial pattern and put $\Sigma_0 = x_1 + \cdots + x_i$. Define $\Sigma_n = \Sigma_{n-1} - W_n X_n$, $L_0 = k$, and $L_n = L_{n-1} - (3X_n + 1)/2$. Then τ is the smallest n such that $L_n \le 0$, and τ is by the strong law finite with probability 1 if $E[3X_n + 1] = 6(p - \frac{1}{3}) > 0$. For $n \le \tau$, Σ_n is the sum of the pattern used to determine W_{n+1}. Since $F_n - F_{n-1} = \Sigma_{n-1} - \Sigma_n$, it follows that $F_n = F_0 + \Sigma_0 - \Sigma_n$ and $F_\tau = F_0 + \Sigma_0$.

7.9. Observe that $E[F_n - F_\tau] = E[\Sigma_{k=1}^n X_k I_{[\tau < k]}] = \Sigma_{k=1}^n E[X_k] P[\tau < k]$.

Section 8

8.8. (b) With probability 1 the population either dies out or goes to infinity. If, for example, $p_{k0} = 1 - p_{k,k+1} = 1/k^2$, then extinction and explosion each have positive probability.

8.9. To prove that $x_i \equiv 0$ is the only possibility in the persistent case, use Problem 8.5, or else argue directly: If $x_i = \Sigma_{j \ne i_0} p_{ij} x_j$, $i \ne i_0$, and K bounds the $|x_i|$, then $x_i = \Sigma p_{ij_1} \cdots p_{j_{n-1}j_n} x_{j_n}$, where the sum is over $j_1, \ldots, j_{n-1}$ distinct from i_0, and hence $|x_i| \le KP_i[X_k \ne i_0, \ k \le n] \to 0$.

8.13. Let P be the set of i for which $\pi_i > 0$, let N be the set of i for which $\pi_i \le 0$, and suppose that P and N are both nonempty. For $i_0 \in P$ and $j_0 \in N$ choose n so that $p_{i_0 j_0}^{(n)} > 0$. Then

$$0 < \sum_{j \in N} \sum_{i \in P} \pi_i p_{ij}^{(n)} = \sum_{j \in N} \pi_j - \sum_{j \in N} \sum_{i \in N} \pi_i p_{ij}^{(n)}$$

$$= \sum_{i \in N} \pi_i \sum_{j \in P} p_{ij}^{(n)} \le 0.$$

Transfer from N to P any i for which $\pi_i = 0$ and use a similar argument.

8.16. Denote the sets (8.32) and (8.52) by P and by F, respectively. Since $F \subset P$, gcd $P \le$ gcd F. The reverse inequality follows from the fact that each integer in P is a sum of integers in F.

8.17. Consider the chain with states $0, 1, \ldots$ and α and transition probabilities $p_{0j} = f_{j+1}$ for $j \ge 0$, $p_{0\alpha} = 1 - f$, $p_{i,i-1} = 1$ for $i \ge 1$, and $p_{\alpha\alpha} = 1$ (α is an absorbing state). The transition matrix is

$$\begin{bmatrix} f_1 & f_2 & f_3 & \cdots & 1-f \\ 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

Show that $f_{00}^{(n)} = f_n$ and $p_{00}^{(n)} = u_n$, and apply Theorem 8.1. Then assume $f = 1$, discard the state α and any states j such that $f_k = 0$ for $k > j$, and apply Theorem 8.8.

In Feller, Volume 1, the renewal theorem is proved by purely analytic means and is then used as the starting point for the theory of Markov chains. Here the procedure is the reverse.

8.19. The transition probabilities are $p_{0r} = 1$ and $p_{i,r-i+1} = p$, $p_{i,r-i} = q$, $1 \le i \le r$; the stationary probabilities are $u_1 = \cdots = u_r = q^{-1}u_0 = (r+q)^{-1}$. The chance of getting wet is $u_0 p$, of which the maximum is $2r + 1 - 2\sqrt{r(r+1)}$. For $r = 5$ this is .046, the pessimal value of p being .523. Of course, $u_0 p \le 1/4r$. In more reasonable climates fewer umbrellas suffice: if $p = .25$ and $r = 3$, then $u_0 p = .050$; if $p = .1$ and $r = 2$, then $u_0 p = .031$. At the other end of the scale, if $p = .8$ and $r = 3$, then $u_0 p = .050$; and if $p = .9$ and $r = 2$, then $u_0 p = .043$.

8.22. For the last part, consider the chain with state space C_m and transition probabilities p_{ij} for $i, j \in C_m$ (show that they do add to 1).

8.23. Let $C' = S - (T \cup C)$, and take $U = T \cup C'$ in (8.51). The probability of absorption in C is the probability of ever entering it, and for initial states i in $T \cup C$ these probabilities are the minimal solution of

$$y_i = \sum_{j \in T} p_{ij} y_j + \sum_{j \in C} p_{ij} + \sum_{j \in C'} p_{ij} y_j, \quad i \in T \cup C',$$

$$0 \le y_i \le 1, \qquad\qquad\qquad\qquad i \in T \cup C'.$$

Since the states in C' ($C' = \varnothing$ is possible) are persistent and C is closed, it is impossible to move from C' to C. Therefore, in the minimal solution of the system above, $y_i = 0$ for $i \in C'$. This gives the system (8.55). It also gives, for the minimal solution, $\sum_{j \in T} p_{ij} y_j + \sum_{j \in C} p_{ij} = 0$, $i \in C'$. This makes probabilistic sense: for an i in C', not only is it impossible to move to a j in C, it is impossible to move to a j in T for which there is positive probability of absorption in C.

8.24. Fix on a state i, and let S_ν consist of those j for which $p_{ij}^{(n)} > 0$ for some n congruent to ν modulo t. Choose k so that $p_{ji}^{(k)} > 0$; if $p_{ij}^{(m)}$ and $p_{ij}^{(n)}$ are positive, then t divides $m + k$ and $n + k$, so that m and n are congruent modulo t. The S_ν are thus well defined.

8.25. Show that Theorem 8.6 applies to the chain with transition probabilities $p_{ij}^{(t)}$.

8.27. (a) From $PC = C\Lambda$ follows $Pc_i = \lambda_i c_i$, from $RP = \Lambda R$ follows $r_i P = \lambda_i r_i$, and from $RC = I$ follows $r_i c_j = \delta_{ij}$. Clearly Λ^n is diagonal and $P^n = C\Lambda^n R$. Hence $p_{ij}^{(n)} = \sum_{u\iota} C_{iu} \lambda_u^n \delta_{u\iota} R_{\iota j} = \sum_u \lambda_u^n (c_u r_u)_{ij} = \sum_u \lambda_u^n (A_u)_{ij}$.

(b) By Problem 8.26, there are scalars ρ and γ such that $r_1 = \rho r_0 = \rho(\pi_1, \ldots, \pi_s)$ and $c_1 = \gamma c_0$, where c_0 is the column vector of 1's. From $r_1 c_1 = 1$ follows $\rho\gamma = 1$, and hence $A_1 = c_1 r_1 = c_0 r_0$ has rows $(\pi_1, \ldots, \pi_s)$. Of course, (8.56) gives the exact rate of convergence. It is useful for numerical work; see Çınlar, pp. 364 ff.

(c) Suppose all four p_{ij} are positive. Then $\pi_1 = p_{21}/(p_{21} + p_{12})$, $\pi_2 = p_{12}/(p_{21} + p_{12})$, the second eigenvalue is $\lambda = 1 - p_{12} - p_{21}$, and

$$P^n = \begin{bmatrix} \pi_1 & \pi_2 \\ \pi_1 & \pi_2 \end{bmatrix} + \lambda^n \begin{bmatrix} \pi_2 & -\pi_2 \\ -\pi_1 & \pi_1 \end{bmatrix}.$$

Note that for given n and ϵ, $\lambda^r > 1 - \epsilon$ is possible, which means that the $p_{ij}^{(n)}$ are not yet near the π_j. In the case of positive p_{ij}, P is always diagonalizable by

$$C = \begin{bmatrix} 1 & \pi_2 \\ 1 & -\pi_1 \end{bmatrix}, \qquad R = \begin{bmatrix} \pi_1 & \pi_2 \\ 1 & -1 \end{bmatrix}$$

(d) For example, take $t = \frac{1}{3}$, $0 < \epsilon < t$, and

$$P = \begin{bmatrix} t & t & t \\ t & t & t \\ t - \epsilon & t + \epsilon & t \end{bmatrix}.$$

In this case, 0 is an eigenvalue with algebraic multiplicity 2 and geometric multiplicity 1.

8.30. Show that $\alpha_n = \pi_i$ and

$$\beta_n - \alpha_n = \frac{2}{n(n-1)} \sum_{k=1}^{n-1} \pi_i (n-k)\left(p_{ii}^{(k)} - \pi_i \right)$$

$$= O\!\left(\frac{1}{n^2} \sum_{k=1}^{n} (n-k)\rho^k \right) = O\!\left(\frac{1}{n} \right),$$

where ρ is as in Theorem 8.9.

8.36. The definitions give

$$E_i\big[f(X_{\sigma_n}) \big] = P_i[\sigma_n < n] f(0) + P_i[\sigma_n = n] f(i + n)$$

$$= 1 - P_i[\sigma_n = n] + P_i[\sigma_n = n](1 - f_{i+n,0})$$

$$= 1 - p_i \cdots p_{i+n-1} + p_1 \cdots p_{i+n-1}(p_{i+n} p_{i+n+1} \cdots),$$

and this goes to 1. Since $P_i[\tau \le n = \sigma_n] \ge P_i([\tau \le n] \cap [X_k > 0,\ k \ge 1]) \to 1 - f_{i0} > 0$, there is an n of the kind required in the last part of the problem. And now

$$E_i[f(X_\tau)] \le P_i[\tau < n = \sigma_n] f(i + n) + 1 - P_i[\tau < n = \sigma_n]$$

$$= 1 - P_i[\tau < n = \sigma_n] f_{i+n,0}.$$

8.37. If $i \ge 1$, $n_1 < n_2$, $(i, \ldots, i + n_1) \in I_{n_1}$, and $(i, \ldots, i + n_2) \in I_{n_2}$, then $P_i[\tau = n_1, \tau = n_2] \ge P_i[X_k = i + k,\ k \le n_2] > 0$, which is impossible.

Section 9

9.3. See Bahadur.

9.7. Because of Theorem 9.6 there are for $P[M_n \ge \alpha]$ bounds of the same order as the ones for $P[S_n \ge \alpha]$ used in the proof of (9.36).

Section 10

10.7. Let μ_1 be counting measure on the σ-field of all subsets of a countably infinite Ω, let $\mu_2 = 2\mu_1$, and let $\mathscr{P}$ consist of the cofinite sets. Granted the existence of Lebesgue measure λ on $\mathscr{R}^1$, one can construct another example: let $\mu_1 = \lambda$ and $\mu_2 = 2\lambda$, and let $\mathscr{P}$ consist of the half-infinite intervals $(-\infty, x]$.

There are similar examples with a field $\mathscr{F}_0$ in place of $\mathscr{P}$. Let Ω consist of the rationals in $(0,1]$, let μ_1 be counting measure, let $\mu_2 = 2\mu_1$, and let $\mathscr{F}_0$ consist of finite disjoint unions of "intervals" $[r \in \Omega: a < r \le b]$.

Section 11

11.4. **(b)** If $(f, g] \subset \bigcup_k (f_k, g_k]$, then $(f(\omega), g(\omega)] \subset \bigcup_k (f_k(\omega), f_k(\omega)]$ for all ω, and Theorem 1.3 gives $g(\omega) - f(\omega) \le \Sigma_k (g_k(\omega) - f_k(\omega))$. If $h_m = (g - f - \Sigma_{k \le n}(g_k - f_k)) \vee 0$, then $h_n \downarrow 0$ and $g - f \le \Sigma_{k \le n}(g_k - f_k) + h_n$. The positivity and continuity of Λ now give $\nu_0(f, g] \le \Sigma_k \nu_0(f_k, g_k]$. A similar, easier argument shows that $\Sigma_k \nu_0(f_k, g_k] \le \nu_0(f, g]$ if $(f_k, g_k]$ are disjoint subsets of $(f, g]$.

11.5. **(b)** From (11.7) it follows that $[f > 1] \in \mathscr{F}_0$ for f in $\mathscr{L}$ Since $\mathscr{L}$ is linear, $[f > x]$ and $[f < -x]$ are in $\mathscr{F}_0$ for $f \in \mathscr{L}$ and $x > 0$. Since the sets (x, ∞) and $(-\infty, -x)$ for $x > 0$ generate $\mathscr{R}^1$, each f in $\mathscr{L}$ is measurable $\sigma(\mathscr{F}_0)$. Hence $\mathscr{F} = \sigma(\mathscr{F}_0)$.

It is easy to show that $\mathscr{F}_0$ is a semiring and is in fact closed under the formation of proper differences. It can happen that $\Omega \notin \mathscr{F}_0$—for example, in the case where $\Omega = \{1, 2\}$ and $\mathscr{L}$ consists of the f with $f(1) = 0$. See Jürgen Kindler: A simple proof of the Daniell–Stone representation theorem. *Amer. Math. Monthly*, **90** (1983), 396–397.)

Section 12

12.4. **(a)** If $\theta_n = \theta_m$, then $\theta_{n-m} = 0$ and $n = m$ because θ is irrational. Split G into finitely many intervals of length less than ϵ; one of them must contain points θ_{2n} and θ_{2m} with $\theta_{2n} < \theta_{2m}$. If $k = m - n$, then $0 < \theta_{2m} - \theta_{2n} = \theta_{2m} \ominus \theta_{2n} = \theta_{2k} < \epsilon$, and the points θ_{2kl} for $1 \le l \le \lfloor \theta_{2k}^{-1} \rfloor$ form a chain in which the distance from each to the next is less than ϵ, the first is to the left of ϵ, and the last is to the right of $1 - \epsilon$.

(c) If $s_1 \ominus s_2 = \theta_{2k+1} \ominus \theta_{2n_1} \oplus \theta_{2n_2}$ lies in the subgroup, then $s_1 = s_2$ and $\theta_{2k+1} = \theta_{2(n_1 - n_2)}$.

12.5. **(a)** The $S \oplus \theta_m$ are disjoint, and $(2n + 1)v + k = (2n + 1)v' + k'$ with $|k|, |k'| \le n$ is impossible if $v \ne v'$.

(b) The $A \oplus \theta_{(2n+1)l}$ are disjoint, contained in G, and have the same Lebesgue measure.

12.6. See Example 2.10 (which applies to any finite measure).

12.8. By Theorem 12.3 and Problem 2.19(b), A contains two disjoint compact sets of arbitrarily small positive measure. Construct inductively compact sets $K_{u_1 \ldots u_n}$ (each u_i is 0 or 1) such that $0 < \mu(K_{u_1 \ldots u_n}) < 3^{-n}$ and $K_{u_1 \ldots u_n 0}$ and $K_{u_1 \ldots u_n 1}$ are disjoint subsets of $K_{u_1 \ldots u_n}$. Take $K = \bigcap_n \bigcup_{u_1 \ldots u_n} K_{u_1 \ldots u_n}$. The Cantor set is a special case.

Section 13

13.3. If $f = \sum_i x_i I_{A_i}$ and $A_i \in T^{-1}\mathscr{F}'$, take A_i' in $\mathscr{F}'$ so that $A_i = T^{-1}A_i'$, and set $\varphi = \sum_i x_i I_{A_i'}$. For the general f measurable $T^{-1}\mathscr{F}'$, there exist simple functions f_n, measurable $T^{-1}\mathscr{F}'$, such that $f_n(\omega) \to f(\omega)$ for each ω. Choose φ_n, measurable $\mathscr{F}'$, so that $f_n = \varphi_n T$. Let C' be the set of ω' for which $\varphi_n(\omega')$ has a finite limit, and define $\varphi(\omega') = \lim_n \varphi_n(\omega')$ for $\omega' \in C'$ and $\varphi(\omega') = 0$ for $\omega' \notin C'$. Theorem 20.1(ii) is a special case.

13.7. The class of Borel functions contains the continuous functions and is closed under pointwise passages to the limit and hence contains $\mathscr{X}$.

By imitating the proof of the π–λ theorem, show that, if f and g lie in $\mathscr{X}$, then so do $f + g$, fg, $f - g$, $f \vee g$ (note that, for example, $[g: f + g \in \mathscr{X}]$ is closed under passages to the limit). If $f_n(x)$ is 1 or $1 - n(x - \alpha)$ or 0 as $x \le \alpha$ or $\alpha \le x \le \alpha + n^{-1}$ or $\alpha + n^{-1} \le x$, then f_n is continuous and $f_n(x) \to I_{(-\infty, \alpha]}(x)$. Show that $[A: I_A \in \mathscr{X}]$ is a λ-system. Conclude that $\mathscr{X}$ contains all indicators of Borel sets, all simple Borel functions, all Borel functions.

13.13. Let $B = \{b_1, \dots, b_k\}$, where $k \le n$, $E_i = C - b_i^{-1}A$, and $E = \bigcup_{i=1}^k E_i$. Then $E = C - \bigcup_{i=1}^k b_i^{-1}A$. Since μ is invariant under rotations, $\mu(E_i) = 1 - \mu(A) < n^{-1}$, and hence $\mu(E) < 1$. Therefore $C - E = \bigcap_{i=1}^k b_i^{-1}A$ is nonempty. Use any θ in $C - E$.

Section 14

14.3. **(b)** Since $u \le F(x)$ is equivalent to $\varphi(u) \le x$, it follows that $u \le F(\varphi(u))$. And since $F(x) < u$ is equivalent to $x < \varphi(u)$, it follows further that $F(\varphi(u) - \epsilon) < u$ for positive ϵ.

14.4. **(a)** If $0 < u < v < 1$, then $P[u \le F(X) < v, \, X \in C] = P[\varphi(u) \le X < \varphi(v), \, X \in C]$. If $\varphi(u) \in C$, this is at most $P[\varphi(u) \le X < \varphi(v)] = F(\varphi(v) -) - F(\varphi(u) -) = F(\varphi(v) -) - F(\varphi(u)) \le v - u$; if $\varphi(u) \notin C$, it is at most $P[\varphi(u) < X < \varphi(v)] = F(\varphi(v) -) - F(\varphi(u)) \le v - u$. Thus $P[F(X) \in [u, v), \, X \in C] \le \lambda[u, v)$ if $0 < u < v < 1$. This is true also for $u = 0$ (let $u \downarrow 0$ and note that $P[F(X) = 0] = 0$) and for $v = 1$ (let $v \uparrow 1$). The finite disjoint unions of intervals $[u, v)$ in $[0, 1)$ form a field there, and by addition $P[F(X) \in A, \, X \in C] \le \lambda(A)$ for A in this field. By the monotone class theorem, the inequality holds for all Borel sets in $[0, 1)$. Since $P[F(X) = 1, \, X \in C] = 0$, this holds also for $A = \{1\}$.

14.5. The sufficiency is easy. To prove necessity, choose continuity points x_i of F in such a way that $x_0 < x_1 < \cdots < x_k$, $F(x_0) < \epsilon$, $F(x_k) > 1 - \epsilon$, and $x_i - x_{i-1} < \epsilon$. If n exceeds some n_0, $|F(x_i) - F_n(x_i)| < \epsilon/2$ for all i. Suppose that $x_{i-1} \le x \le x_i$. Then $F_n(x) \le F_n(x_i) \le F(x_i) + \epsilon/2 \le F(x + \epsilon) + \epsilon/2$. Establish a similar inequality going the other direction, and give special arguments for the cases $x \le x_0$ and $x \ge x_k$.

Section 15

15.1. Suppose there is an $\mathscr{F}$-partition such that $\sum_i [\sup_{A_i} f] \mu(A_i) < \infty$. Then $a_i = \sup_{A_i} f < \infty$ for i in the set I of indices for which $\mu(A_i) > 0$. If $a = \max_I a_i$, then $\mu[f > a] = \sum_I \mu(A_i \cap [f > a]) \le \sum_I \mu(A_i \cap [f > a_i]) = 0$. And $A_i \cap [f > 0] = \varnothing$ for i outside the set J of indices for which $\mu(A_i) < \infty$, so that $\mu[f > 0] = \sum_I \mu(A_i \cap [f > 0]) \le \sum_J \mu(A_i) < \infty$.

15.4. Let $(\Omega, \mathscr{F}^+, \mu^+)$ be the completion (Problems 3.10 and 10.5) of $(\Omega, \mathscr{F}, \mu)$. If g is measurable $\mathscr{F}$, $[f \neq g] \subset A$, $A \in \mathscr{F}$, $\mu(A) = 0$, and $H \in \mathscr{R}^1$, then $[f \in H] = (A^c \cap [f \in H]) \cup (A \cap [f \in H]) = (A^c \cap [g \in H]) \cup (A \cap [f \in H])$ lies in $\mathscr{F}^+$, and hence f is measurable $\mathscr{F}^+$.

(a) Since f is measurable $\mathscr{F}^+$, it will be enough to prove that for each (finite) $\mathscr{F}^+$-partition $\{B_j\}$ there is an $\mathscr{F}$-partition $\{A_i\}$ such that $\Sigma_i[\inf_{A_i} f]\mu(A_i) \geq \Sigma_j[\inf_{B_j} f]\mu^+(B_j)$, and to prove the dual relation for the upper sums. Choose (Problem 3.2) $\mathscr{F}$-sets A_i so that $A_i \subset B_i$ and $\mu(A_i) = \mu_*(B_i) = \mu^+(B_i)$. For the partition consisting of the A_i together with $(\cup_i A_i)^c$, the lower sum is at least $\Sigma_i[\inf_{B_i} f]\mu(A_i) = \Sigma_i[\inf_{B_i} f]\mu^+(B_i)$.

(b) Choose successively finer $\mathscr{F}$-partitions $\{A_{ni}\}$ in such a way that the corresponding upper and lower sums differ by at most $1/n^3$. Let g_n and f_n have values $\inf_{A_{ni}} f$ and $\sup_{A_{ni}} f$ on A_{ni}. Use Markov's inequality—since $\mu(\Omega)$ is finite, it may as well be 1—to show that $\mu[f_n - g_n \geq 1/n] \leq 1/n^2$, and then use the first Borel-Cantelli lemma to show that $f_n - g_n \to 0$ almost everywhere. Take $g = \lim_n g_n$.

Section 16

16.3. $0 \leq f_n - f_1 \uparrow f - f_1$.

16.4. (a) By Fatou's lemma,

$$\int f \, d\mu - \int a \, d\mu = \int \lim_n (f_n - a_n) \, d\mu$$

$$\leq \lim_n \inf \int (f_n - a_n) \, d\mu = \lim_n \inf \int f_n \, d\mu - \int a \, d\mu$$

and

$$\int b \, d\mu - \int f \, d\mu = \int \lim_n (b_n - f_n) \, d\mu$$

$$\leq \lim_n \inf \int (b_n - f_n) \, d\mu = \int b \, d\mu - \lim_n \sup \int f_n \, d\mu.$$

Therefore

$$\lim_n \sup \int f_n \, d\mu \leq \int f \, d\mu \leq \lim_n \inf \int f_n \, d\mu.$$

16.6. For $\omega \in A$ and small enough complex h,

$$|f(\omega, z_0 + h) - f(\omega, z_0)| = \left| \int_{z_0}^{z_0 + h} f'(\omega, z) \, dz \right| \leq |h| g(\omega, z_0).$$

16.8. Use the fact that $\int_A |f| \, d\mu \leq \alpha \mu(A) + \int_{[|f| \geq \alpha]} |f| \, d\mu$.

16.9. If $\mu(A) < \delta$ implies $\int_A |f_n| d\mu < \epsilon$ for all n, and if $\alpha^{-1} \sup_n \int |f_n| d\mu < \delta$, then $\mu[|f_n| \geq \alpha] \leq \alpha^{-1} \int |f_n| d\mu < \delta$ and hence $\int_{[|f_n| \geq \alpha]} |f_n| d\mu < \epsilon$ for all n. For the reverse implication adapt the argument in the preceding note.

16.10. **(b)** Suppose that f_n are nonnegative and satisfy condition (ii) and μ is nonatomic. Choose δ so that $\mu(A) \leq \delta$ implies $\int_A f_n d\mu \leq 1$ for all n. If $\mu[f_n = \infty] > 0$, there is an A such that $A \subset [f_n = \infty]$ and $0 < \mu(A) < \delta$; but then $\int_A f_n d\mu = \infty$. Since $\mu[f_n = \infty] = 0$, there is an α such that $\mu[f_n > \alpha] \leq \delta \leq \mu[f_n \geq \alpha]$. Choose $B \subset [f_n = \alpha]$ in such a way that $A = [f_n > \alpha] \cup B$ satisfies $\mu(A) = \delta$. Then $\alpha\delta = \alpha\mu(A) \leq \int_A f_n d\mu \leq 1$ and $\int f_n d\mu \leq 1 + \alpha\mu(A^c) \leq 1 + \delta^{-1}\mu(\Omega)$.

16.12. **(b)** Suppose that $f \in \mathscr{L}$ and $f \geq 0$. If $f_n = (1 - n^{-1})f \vee 0$, then $f_n \in \mathscr{L}$ and $f_n \uparrow f$, so that $\nu(f_n, f] = \Lambda(f - f_n) \downarrow 0$. Since $\nu(f_1, f] < \infty$, it follows that $\nu[(\omega, t): f(\omega) = t] = 0$. The disjoint union

$$B_n = \bigcup_{i=1}^{n2^n} \left(\left[\frac{i}{2^n} < f \leq \frac{i+1}{2^n} \right] \times \left(0, \frac{i}{2^n} \right] \right)$$

increases to B, where $B \subset (0, f]$ and $(0, f] - B \subset [(\omega, t): f(\omega) = t]$. Therefore

$$\Lambda(f) = \nu(0, f] = \lim_n \nu(B_n) = \lim_n \sum_{i=1}^{n2^n} \frac{i}{2^n} \mu\left[\frac{i}{2^n} < f \leq \frac{i+1}{2^n} \right] = \int f d\mu.$$

Section 17

17.1. **(a)** Let A_ϵ be the set of x such that for every δ there are points y and z satisfying $|y - x| < \delta$, $|z - x| < \delta$, and $|f(y) - f(z)| \geq \epsilon$. Show that A_ϵ is closed and D_f is the union of the A_ϵ.

(c) Given ϵ and η, choose a partition into intervals I_i for which the corresponding upper and lower sums differ by at most $\epsilon\eta$. By considering those I_i whose interiors meet A_ϵ, show that $\epsilon\eta \geq \epsilon\lambda(A_\epsilon)$.

(d) Let M bound $|f|$ and, given ϵ, find an open G such that $D_f \subset G$ and $\lambda(G) < \epsilon/M$. Take $C = [0, 1] - G$ and show by compactness that there is a δ such that $|f(y) - f(x)| < \epsilon$ if x (but perhaps not y) lies in C and $|y - x| < \delta$. If $[0, 1]$ is decomposed into intervals I_i with $\lambda(I_i) < \delta$, and if $x_i \in I_i$, let g be the function with value $f(x_i)$ on I_i. Let Σ' denote summation over those i for which I_i meets C, and let Σ'' denote summation over the other i. Show that

$$\left| \int_0^1 f(x) \, dx - \Sigma f(x_i)\lambda(I_i) \right| \leq \int_0^1 |f(x) - g(x)| \, dx$$

$$\leq \Sigma' 2\epsilon\lambda(I_i) + \Sigma'' 2M\lambda(I_i) < 4\epsilon.$$

17.10. **(c)** Do not overlook the possibility that points in $(0, 1) - K$ converge to a point in K.

17.11. **(b)** Apply the bounded convergence theorem to $f_n(x) = (1 - n \operatorname{dist}(x, [s, t]))^+$.

(c) The class of Borel sets B in $[u, v]$ for which $f = I_B$ satisfies (17.8) is a λ-system.

(e) Choose simple f_n such that $0 \le f_n \uparrow f$ To (17.8) for $f = f_n$, apply the monotone convergence theorem on the right and the dominated convergence theorem on the left.

17.12. If $g(x)$ is the distance from x to $[a, b]$, then $f_n = (1 - ng) \vee 0 \downarrow I_{[a,b]}$ and $f_n \in \mathcal{L}$; since the continuous functions are measurable $\mathcal{R}^1$, it follows that $\mathcal{F} = \mathcal{R}^1$. If $f_n(x) \downarrow 0$ for each x, then the compact sets $[x: f_n(x) \ge \epsilon]$ decrease to $\varnothing$ and hence one of them is $\varnothing$; thus the convergence is uniform.

17.13. The linearity and positivity of Λ are certainly elementary facts, and for the continuity property, note that if $0 \le f \le \epsilon$ and f vanishes outside $[a, b]$, then elementary considerations show that $0 \le \Lambda(f) \le \epsilon(b - a)$.

Section 18

18.2. First, $\mathcal{X} \times \mathcal{X}$ is generated by the sets of the forms $\{x\} \times X$ and $X \times \{x\}$. If the diagonal E lies in $\mathcal{X} \times \mathcal{X}$, then there must be a countable S in X such that E lies in the σ-field $\mathcal{F}$ generated by the sets of these two forms for x in S. If $\mathcal{P}$ consists of S^c and the singletons in S, then $\mathcal{F}$ is the class of unions of sets in the partition $[P_1 \times P_2: P_1, P_2 \in \mathcal{P}]$. But $E \in \mathcal{F}$ is impossible.

18.3. Consider $A \times B$, where A consists of a single point and B lies outside the completion of $\mathcal{R}^1$ with respect to λ.

18.17. Put $f_p = p^{-1} \log p$, and put $f_n = 0$ if n is not a prime In the notation of (18.17), $F(x) = \log x + \varphi(x)$, where φ is bounded because of (5.51). If $G(x) = -1/\log x$, then

$$\sum_{p \le x} \frac{1}{p} = \frac{F(x)}{\log x} + \int_2^x \frac{F(t)\, dt}{t \log^2 t}$$

$$= 1 + \frac{\varphi(x)}{\log x} + \int_2^x \frac{dt}{t \log t} + \int_2^\infty \frac{\varphi(t)\, dt}{t \log^2 t} - \int_x^\infty \frac{\varphi(t)\, dt}{t \log^2 t}.$$

Section 19

19.3. See BANACH, p. 34.

19.4. **(a)** Take $f = 0$ and $f_n = I_{(0, 1/n)}$.
(b) Take $f = 0$, and let $\{f_n\}$ be an infinite orthonormal set. Use the fact that $\sum_n (f_n, g)^2 \le \|g\|^2$.

19.5. Take $f_n = n I_{(0, 1/n)}$, and suppose that f_{n_k} converges weakly to some f in L^1. Integrate against the L^∞-functions $\operatorname{sgn} f \cdot I_{(\epsilon, 1)}$ and conclude that $f = 0$ almost everywhere; now integrate against the function identically 1 and get a contradiction.

Section 20

20.4. Suppose $U_1, \ldots, U_k$ are independent and uniformly distributed over the unit interval, put $V_i = 2nU_i - n$, and let μ_n be $(2n)^k$ times the distribution of

$(V_1, \ldots, V_k)$. Then μ_n is supported by $Q_n = (-n, n] \times \cdots \times (-n, n]$, and if $I = (a_1, b_1] \times \cdots \times (a_k, b_k] \subset Q_n$, then $\mu_n(I) = \prod_{i=1}^{k} (b_i - a_i)$. Further, if $A \subset Q_n \subset Q_m$ $(n < m)$, then $\mu_n(A) = \mu_m(A)$. Define $\lambda_k(A) = \lim_n \mu_n(A \cap Q_n)$.

20.7. By the argument preceding (8.16),

$$P_i[T_1 = n_1, \ldots, T_k = n_k] = f_{ij}^{(n_1)} f_{jj}^{(n_2 - n_1)} \cdots f_{jj}^{(n_k - n_{k-1})}.$$

For the general initial distribution, average over i.

20.8. (a) Use the π–λ theorem to show that $P[(X_{\pi 1}, \ldots, X_{\pi n}) \in H]$ is the same for all permutations π.

(b) Use part (a) and the fact that $Y_n = r$ if and only if $T_r^{(n)} = n$.

(c) If $k \le n$, then $Y_k = r$ if and only if exactly $r - 1$ among the integers $1, \ldots, k - 1$ precede k in the permutation $T^{(n)}$.

(d) Observe that $T^{(n)} = (t_1, \ldots, t_n)$ and $Y_{n+1} = r$ if and only if $T^{(n+1)} = (t_1, \ldots, t_{r-1}, n+1, t_r, \ldots, t_n)$, and conclude that $\sigma(Y_{n+1})$ is independent of $\sigma(T^{(n)})$ and hence of $\sigma(Y_1, \ldots, Y_n)$—see Problem 20.6.

20.12. If X and Y are independent, then

$$P\big[|(X + Y) - (x + y)| < \epsilon\big] \ge P\big[|X - x| < \tfrac{1}{2}\epsilon\big] P\big[|Y - y| < \tfrac{1}{2}\epsilon\big]$$

and

$$P[X + Y = x + y] \ge P[X = x]P[Y = y].$$

20.14. The partial-fraction expansion gives

$$c_u(y - x)c_t(x) = \frac{uv}{\pi^2} \frac{1}{R}(A + B + C + D),$$

where $R = (u^2 - v^2)^2 + 2(u^2 + v^2)y^2 + y^4$ and

$$A = \frac{y^2 + v^2 - u^2}{u^2 + (y - x)^2}, \qquad B = \frac{2y(y - x)}{u^2 + (y - x)^2},$$

$$C = \frac{y^2 - v^2 + u^2}{v^2 + x^2}, \qquad D = \frac{2yx}{v^2 + x^2}.$$

After the fact this can of course be checked mechanically. Integrate over $[-t, t]$ and let $t \to \infty$: $\int_{-t}^{t} D \, dx = 0$, $\int_{-t}^{t} B \, dx \to 0$, and $\int_{-\infty}^{\infty} (A + C) \, dx = (y^2 + v^2 - u^2)u^{-1}\pi + (y^2 - v^2 + u^2)v^{-1}\pi = u^{-1}v^{-1}\pi^2 R c_{u+t}(y)$. There is a very simple proof by characteristic functions; see Problem 26.9.

20.16. See Example 20.1 for the case $n = 1$, prove by inductive convolution and a change of variable that the density must have the form $K_n x^{(n/2)-1} e^{-x/2}$, and then from the fact that the density must integrate to 1 deduce the form of K_n.

20.17. Show by (20.38) and a change of variable that the left side of (20.48) is some constant times the right side; then show that the constant must be 1.

20.20. (a) Given ϵ choose M so that $P[|X| > M] < \epsilon$ and $P[|Y| > M] < \epsilon$, and then choose δ so that $|x|, |y| \le M$, $|x - x'| < \delta$, and $|y - y'| < \delta$ imply that $|f(x', y') - f(x, y)| < \epsilon$. Note that $P[|f(X_n, Y_n) - f(X, Y)| \ge \epsilon] \le 2\epsilon + P[|X_n - X| \ge \delta] + P[|Y_n - Y| \ge \delta]$.

20.23. Take, for example, independent X_n assuming the values 0 and n with probabilities $1 - n^{-1}$ and n^{-1}. Estimate the probability that $X_k = k$ for some k in the range $n/2 < k \le n$.

20.24. (b) For each m split A into 2^m sets A_{mk} of probability $P(A)/2^M$. Arrange all the A_{mk} in one infinite sequence, and let X_n be the indicator of the nth set in it.

20.27. To get the distribution of Φ, show by integration that for $0 \le \phi \le 2\pi$, the intersection with the unit ball of the (x_1, x_2, x_3)-set where $0 \le x_3 \le (x_1^2 + x_2^2)^{1/2} \tan \phi$ has volume $\frac{2}{3}\pi \sin \phi$.

Section 21

21.5. Consider ΣI_{A_n}. A random variable is finite with probability 1 if (but not only if) it is integrable.

21.6. Calculate $\int_0^\infty x \, dF(x) = \int_0^\infty \int_0^x dy \, dF(x) = \int_0^\infty \int_y^\infty dF(x) \, dy$.

21.8. (a) Write $E[Y - X] = \int_{X < Y} \int_{X < t \le Y} dt \, dP - \int_{Y < X} \int_{Y < t \le X} dt \, dP$.

21.10. (a) The most important dependent uncorrelated random variables are the trigonometric functions—the random variables $\sin 2\pi n\omega$ and $\cos 2\pi n\omega$ on the unit interval with Lebesgue measure. See Problem 19.8.

21.13. Use Fubini's theorem; see (20.29) and (20.30).

21.14. Even if $X = -Y$ is not integrable, $X + Y = 0$ is. Since $|Y| \le |x| + |x + Y|$, $E[|Y|] = \infty$ implies that $E[|x + Y|] = \infty$ for each x; use Problem 21.13. See also the lemma in Section 28.

21.21. Use (21.12).

Section 22

22.2. For sufficiency, use $E[\Sigma |X_n^{(c)}|] = \Sigma E[|X_n^{(c)}|]$.

22.8. (a) Put $U = \Sigma_k I_{[k \le \tau]} X_k^+$ and $V = \Sigma_k I_{[k \le \tau]} X_k^-$, so that $S_\tau = U - V$. Since $[\tau \ge k] = \Omega - [\tau \le k - 1]$ lies in $\sigma(X_1, \ldots, X_{k-1})$, it follows that $E[I_{[\tau \ge k]} X_k^+] = E[I_{[\tau \ge k]}]E[X_k^+] = P[\tau \ge k]E[X_1^+]$. Hence $E[U] = \Sigma_{k=1}^\infty E[X_1^+]P[\tau \ge k] = E[X_1^+]E[\tau]$. Treat V the same way.

(b) To prove $E[\tau] < \infty$, show that $P[\tau > (a + b)n] \le (1 - p^{a+b})^n$. By (7.7), S_τ is b with probability $(1 - \rho^a)/(1 - \rho^{a+b})$ and $-a$ with the opposite probability. Since $E[X_1] = p - q$,

$$E[\tau] = \frac{a}{q - p} - \frac{a + b}{q - p} \frac{1 - \rho^a}{1 - \rho^{a+b}}, \qquad \rho = \frac{q}{p} \ne 1.$$

22.11. For each θ, $\sum_n e^{iX_n}(e^{i\theta}z)^n$ has the same probabilistic behavior as the original series, because the $X_n + n\theta$ reduced modulo 2π are independent and uniformly distributed. Therefore, the rotation idea in the proof of Theorem 22.9 carries over. See KAHANE for further results.

22.14. **(b)** Let $A = f^{-1}B$ and suppose p is a period of f. Let $m = \lfloor x/p \rfloor$ and $n = \lfloor 1/p \rfloor$. By periodicity, $P(A \cap [y, y + p])$ is the same for all y; therefore, $|P(A \cap [0, x]) - mP(A \cap [0, p])| \le p$, $|P(A) - nP(A \cap [0, p])| \le p$, and $|P(A \cap [0, x]) - P(A)x| \le 2p + |x - m/n| \le 3p$. Since p can be taken arbitrarily small, $P(A \cap [0, x]) = P(A)x$.

22.15. **(a)** By the inequalities $L(s) \le M(s)$ and (22.24),

$$B_E(2s) = 1 \wedge 3L(2s) \le 3M(2s) \le 3B_O(s).$$

For the other inequality, note first that $T(s)$ is nonincreasing and that $R(2s) \le 2L(s)$ and $T(s) \le L(s) \le M(s)$. If $B_E(s) \le 1/3$, then

$$B_O(6s) \le \frac{T(6s)}{1 - R(6s)} \le \frac{T(3s)}{1 - 2L(3s)} \le \frac{M(3s)}{1 - 2M(3s)}$$

$$\le \frac{B_E(s)}{1 - 2B_E(s)} \le 3B_E(s). \quad \cdot$$

On the other hand, if $B_E(s) > 1/3$, then $B_O(6s) \le 1 < 3B_E(s)$. In either case, $B_O(6s) \le 3B_E(s)$.

Section 23

23.3. Note that A_t cannot exceed t. If $0 \le u \le t$ and $v \ge 0$, then $P[A_t \ge u, B_t > v] = P[N_{t+v} - N_{t-u} = 0] = e^{-\alpha u}e^{-\alpha v}$.

23.4. **(a)** Use (20.37) and the distributions of A_t and B_t.
(b) A long interarrival interval has a better chance of covering t than a short one does.

23.6. The probability that $N'_{S_{n+k}} - N'_{S_n} = j$ is

$$\int_0^\infty e^{-\beta x} \frac{(\beta x)^j}{j!} \frac{\alpha^k}{\Gamma(k)} x^{k-1}e^{-\alpha x}\,dx = \frac{\alpha^k \beta^j}{(\alpha + \beta)^{k+j}} \frac{(j + k - 1)!}{j!(k - 1)!}.$$

23.8. Let M_t be the given process and put $\varphi(t) = E[M_t]$. Since there are no fixed discontinuities, $\varphi(t)$ is continuous. Let $\psi(u) = \inf[t: u \le \varphi(t)]$, and show that $N_u = M_{\psi(u)}$ is an ordinary Poisson process and $M_t = N_{\varphi(t)}$.

23.9. Let $t \to \infty$ in

$$\frac{S_{N_t}}{N_t} \le \frac{t}{N_t} \le \frac{S_{N_t+1}}{N_t+1} \frac{N_t + 1}{N_t}.$$

23.11. Restrict t in Problem 23.10 to integers. The waiting times are the Z_n of Problem 20.7, and account must be taken of the fact that the distribution of Z_1 may differ from that of the other Z_n.

Section 25

25.1. (e) Let G be an open set that contains the rationals and satisfies $\lambda(G) < \frac{1}{2}$. For $k = 0, 1, \ldots, n - 1$, construct a triangle whose base contains k/n and is contained in G: make these bases so narrow that they do not overlap, and adjust the heights of the triangles so that each has area $1/n$. For the nth density, piece together these triangular functions, and for the limit density, use the function identically 1 over the unit interval.

25.2. By Problem 14.8 it suffices to prove that $F_n(\ ,\omega) \Rightarrow F$ with probability 1, and for this it is enough that $F_n(x, \omega) \to F(x)$ with probability 1 for each rational x.

25.3. (b) It can be shown, for example, that (25.14) holds for $x_n = n!$. See Persi Diaconis: The distribution of leading digits and uniform distribution $\bmod 1$, *Ann. Prob.*, 5 (1977), 72–81.

(c) The first significant digits of numbers drawn at random from empirical compilations such as almanacs and engineering handbooks seem approximately to follow the limiting distribution in (25.15) rather than the uniform distribution over $1, 2, \ldots, 9$. This is sometimes called *Benford's law*. One explanation is that the distribution of the observation X and hence of $\log_{10} X$ will be spread over a large interval; if $\log_{10} X$ has a reasonably smooth density, it then seems plausible that $\{\log_{10} X\}$ should be approximately uniformly distributed. See Feller, Volume 2, p. 62.

25.9. Use Scheffé's theorem.

25.10. Put $f_n(x) = P[X_n = \gamma_n + k\delta_n]\delta_n^{-1}$ for $\gamma_n + k\delta_n < x \leq \gamma_n + (k + 1)\delta_n$. Construct random variables Y_n with densities f_n, and first prove $Y_n \Rightarrow X$. Show that $Z_n = \gamma_n + [(Y_n - \gamma_n)/\delta_n]\delta_n$ has the distribution of X_n and that $Y_n - Z_n \Rightarrow 0$.

25.11. For a proof of (25.16) see Feller, Volume 1, Chapter 7.

25.13. (b) Follow the proof of Theorem 25.8, but approximate $I_{(x, y]}$ instead of $I_{(-\infty, x]}$.

25.20. Let X_n assume the values n and 0 with probabilities $p_n = 1/(n \log n)$ and $1 - p_n$.

Section 26

26.1. (b) Let μ be the distribution of X. If $|\varphi(t)| = 1$ and $t \neq 0$, then $\varphi(t) = e^{ita}$ for some a, and $0 = \int_{-\infty}^{\infty}(1 - e^{it(x-a)})\mu(dx) = \int_{-\infty}^{\infty}(1 - \cos t(x - a))\mu(dx)$. Since the integral vanishes, μ must confine its mass to the points where the nonnegative integrand vanishes, namely to the points x for which $t(x - a) = 2\pi n$ for some integer n.

(c) The mass of μ concentrates at points of the form $a + 2\pi n/t$ and also at points of the form $a' + 2\pi n/t'$. If μ is positive at two distinct points, it follows that t/t' is rational.

26.3. (a) Let $f_0(x) = \pi^{-1}x^{-2}(1 - \cos x)$ be the density corresponding to $\varphi_0(t)$. If $p_k = (s_k - s_{k+1})t_k$, then $\sum_{k=1}^\infty p_k = 1$; since $\sum_{k=1}^\infty p_k \varphi_0(t/t_k) = \varphi(t)$ (check the points $t = t_j$), $\varphi(t)$ is the characteristic function of the continuous density $\sum_{k=1}^\infty p_k t_k f_0(t_k, x)$.

(b) If $\lim_{t \to \infty} \varphi(t) = 0$, approximate φ by functions of the kind in part (a), pass to the limit, and use the first corollary to the continuity theorem. If φ does not vanish at infinity, mix in a unit mass at 0.

26.12. On the right in (26.30) replace $\varphi(t)$ by the integral defining it and apply Fubini's theorem; the integral average comes to

$$\mu\{a\} + \int_{x \neq a} \frac{\sin T(x - a)}{T(x - a)} \mu(dx).$$

Now use the bounded convergence theorem.

26.15. (a) Use (26.4_0) to prove that $|\varphi_n(t + h) - \varphi_n(t)| \leq 2\mu_n(-a. a)^c + a|h|$.

(b) Use part (a).

26.17. (a) Use the second corollary to the continuity theorem.

26.19. For the Weierstrass approximation theorem, see RUDIN$_1$, Theorem 7.32.

26.22. (a) If a_n goes to 0 along a subsequence, then $|\psi(t)| \equiv 1$; use part (c) of Problem 26.1.

(c) Suppose two subsequences of $\{a_n\}$ converge to a_0 and a, where $0 < a_0 < a$; put $\theta = a_0/a$ and show that $|\varphi(t)| = |\varphi(\theta^k t)|$.

(d) Observe that

$$b_n = -i[e^{itb_n} - 1]\left[\int_0^t e^{isb_n} ds\right]^{-1}.$$

26.25. First do the nonnegative case; then note that if f and g have the same coefficients, so do $f^+ + g^-$ and $g^+ + f^-$

Section 27

27.8. By the same reasoning as in Example 27.3, $(R_n - \log n)/\sqrt{\log n} \Rightarrow N$.

27.9. The Lindeberg theorem applies: $(S_n - n^2/4)/\sqrt{n^3/36} \Rightarrow N$.

27.11. Let Y_n be X_n or 0 according as $|X_n| \leq n^{1/2} \log n$ or not. Show that $X_n = Y_n$ for large n, with probability 1, and that Lyapounov's theorem ($\delta = 1$) applies to the Y_n.

27.12. For example, let the distribution of X_n be the mixture, with weights $1 - n^{-2}$ and n^{-2}, of the standard normal and Cauchy distributions.

27.16. Write $\int_x^\infty e^{-u^2/2} du = x^{-1}e^{-x^2/2} - \int_x^\infty u^{-2}e^{-u^2/2} du$.

27.17. For another approach to large-deviation theory, see Mark Pinsky: An elementary derivation of Khintchine's estimate for large deviations, *Proc. Amer. Math. Soc.*, **22** (1969), 288–290.

27.19. (a) Everything comes from (4.7). If $A = [(l_1, \ldots, l_k) \in H]$ and $B \in \sigma(l_{k+n}, l_{k+n+1}, \ldots)$, then

$$|P(A \cap B) - P(A)P(B)|$$

$$\leq \sum |P([l_u = i_u, u \leq k] \cap B) - P[l_u = i_u, u \leq k]P(B)|,$$

where the sum extends over the k-tuples $(i_1, \ldots, i_k)$ of nonnegative integers in H. The summand vanishes if $u + i_u < k + n$ for $u \leq k$, the remaining terms add to at most $2\sum_{u=1}^{k} P[l_u \geq k + n - u] \leq 4/2^n$

(b) To show that $\sigma^2 = 6$ (see (27.20)), show that l_1 has mean 1 and variance 2 and that

$$\int_{[l_1 = i]} l_1 l_{1+n} \, dP = \begin{cases} P[l_1 = i]iE[l_{1+n}] & \text{if } i < n, \\ P[l_1 = i]i(i-n) & \text{if } i \geq n. \end{cases}$$

Section 28

28.2. (b) Pass to a subsequence along which $\mu_n(R^1) \to \infty$, choose ϵ_n so that it decreases to 0 and $\epsilon_n \mu_n(R^1) \to \infty$, and choose x_n so that it increases to ∞ and $\mu_n(-x_n, x_n) > \frac{1}{2}\mu_n(R^1)$; consider the f that satisfies $f(\pm x_n) = \epsilon_n$ for all n and is defined by linear interpolation in between these points.

28.4. (a) If all functions (28.12) are characteristic functions, they are all certainly infinitely divisible. Since (28.12) is continuous at 0, it need only be exhibited as a limit of characteristic functions. If μ_n has density $I_{[-n,n]}(1 + x^2)$ with respect to ν, then

$$\exp\left[i\gamma t + it\int_{-\infty}^{\infty} \frac{x}{1+x^2}\mu_n(dx) + \int_{-\infty}^{\infty}\left(e^{itx} - 1 - itx\right)\frac{1}{x^2}\mu_n(dx)\right]$$

is a characteristic function and converges to (28.12). It can also be shown that every infinitely divisible distribution (no moments required) has characteristic function of the form (28.12); see Gnedenko & Kolmogorov, p. 76.

(b) Use (see Problem 18.19) $-|t| = \pi^{-1}\int_{-\infty}^{\infty}(\cos tx - 1)x^{-2} \, dx$.

28.14. If $X_1, X_2, \ldots$ are independent and have distribution function F, then $(X_1 + \cdots + X_n)/\sqrt{n}$ also has distribution function F. Apply the central limit theorem.

28.15. The characteristic function of Z_n is

$$\exp\frac{c}{n}\sum_k \frac{1}{(|k|/n)^{1+\alpha}}\left(e^{itk/n} - 1\right) \to \exp c\int_{-\infty}^{\infty}\frac{e^{itx} - 1}{|x|^{1+\alpha}} \, dx$$

$$= \exp\left[-c|t|^{\alpha}\int_{-\infty}^{\infty}\frac{1 - \cos x}{|x|^{1+\alpha}} \, dx\right].$$

Section 29

29.1. (a) If f is lower semicontinuous, $[x: f(x) > t]$ is open. If f is positive, which is no restriction, then $\int f \, d\mu = \int_0^\infty \mu[f > t] \, dt \le \int_0^\infty \liminf_n \mu_n[f > t] \, dt \le \liminf_n \int_0^\infty \mu_n[f > t] \, dt = \liminf_n \int f \, d\mu_n$.

(b) If G is open, then I_G is lower semicontinuous.

29.7. Let Σ be the covariance matrix. Let M be an orthogonal matrix such that the entries of $M\Sigma M'$ are 0 except for the first r diagonal entries, which are 1. If $Y = MX$, then Y has covariance matrix $M\Sigma M'$, and so $Y = (Y_1, \ldots, Y_r, 0, \ldots, 0)$, where $Y_1, ., Y_r$ are independent and have the standard normal distribution. But $|X|^2 = \Sigma_{i=1}^k Y_i^2$.

29.8. By Theorem 29.5, X_n has asymptotically the centered normal distribution with covariances σ_{ij}. Put $x = (p_1^{1/2}, \ldots, p_k^{1/2})$ and show that $\Sigma x' = 0$, so that 0 is an eigenvalue of Σ. Show that $\Sigma y' = y'$ if y is perpendicular to x, so that Σ has 1 as an eigenvalue of multiplicity $k - 1$. Use Problem 29.7 together with Theorem 29.2 $(h(x) = |x|^2)$.

29.9. (a) Note that $n^{-1} \Sigma_{i=1}^n Y_{ni}^2 \Rightarrow 1$ and that $(X_{n1}, \ldots, X_{nt})$ has the same distribution as $(Y_{n1}, \ldots, Y_{nt})/(n^{-1} \Sigma_{i=1}^n Y_{ni}^2)^{1/2}$.

Section 30

30.1. Rescale so that $s_n^2 = 1$, and put $L_n(\epsilon) = \Sigma_k \int_{|X_{nk}| \ge \epsilon} X_{nk}^2 \, dP$. Choose increasing n_u so that $L_n(u^{-1}) \le u^{-3}$ for $n \ge n_u$, and put $M_n = u^{-1}$ for $n_u \le n < n_{u+1}$. Then $M_n \to 0$ and $L_n(M_n) \le M_n^3$. Put $Y_{nk} = X_{nk} I_{[|X_{nk}| \le M_n]}$. Show that $\Sigma_k E[Y_{nk}] \to 0$ and $\Sigma_k E[Y_{nk}^2] \to 1$, and apply to $\Sigma_k Y_{nk}$ the central limit theorem under (30.5). Show that $\Sigma_k P[X_{nk} \ne Y_{nk}] \to 0$.

30.4. Suppose that the moment generating function M_n of μ_n converges to the moment generating function M of μ in some interval about s. Let ν_n have density $e^{sx}/M_n(s)$ with respect to μ_n, and let ν have density $e^{sx}/M(s)$ with respect to μ. Then the moment generating function of ν_n converges to that of ν in some interval about 0, and hence $\nu_n \Rightarrow \nu$. Show that $\int_{-\infty}^\infty f(x) \mu_n(dx) \to \int_{-\infty}^\infty f(x) \mu(dx)$ if f is continuous and has bounded support; see Problem 25.13(b).

30.5. (a) By Hölder's inequality $|\Sigma_{j=1}^k t_j x_j|^r \le k^{r-1} \Sigma_{j=1}^k |t_j x_j|^r$, and so $\Sigma_r \theta^r \int |\Sigma_j t_j x_j|^r \mu(dx)/r!$ has positive radius of convergence. Now

$$\int_{R^k} \left(\sum_{j=1}^k t_j x_j \right)^r \mu(dx) = \sum t_1^{r_1} \cdots t_k^{r_k} \alpha(r_1, \ldots, r_k),$$

where the summation extends over k-tuples that add to r. Project μ to the line by the mapping $\Sigma_j t_j x_j$, apply Theorem 30.1, and use the fact that μ is determined by its values on half-spaces.

30.6. Use the Cramér–Wold idea.

30.8. Suppose that $k = 2$ in (30.30). Then

$$M\left[(\cos \lambda_1 x)^{r_1}(\cos \lambda_2 x)^{r_2}\right]$$

$$= M\left[\left(\frac{e^{i\lambda_1 x} + e^{-i\lambda_1 x}}{2}\right)^{r_1}\left(\frac{e^{i\lambda_2 x} + e^{-i\lambda_2 x}}{2}\right)^{r_2}\right]$$

$$= 2^{-r_1 - r_2} \sum_{j_1 = 0}^{r_1} \sum_{j_2 = 0}^{r_2} \binom{r_1}{j_1}\binom{r_2}{j_2} M\left[\exp i\left(\lambda_1(2j_1 - r_1) + \lambda_2(2j_2 - r_2)\right)x\right].$$

By (26.33) and the independence of λ_1 and λ_2, the last mean here is 1 if $2j_1 - r_1 = 2j_2 - r_2 = 0$ and is 0 otherwise. A similar calculation for $k = 1$ gives (30 28), and a similar calculation for general k gives (30.30). The actual form of the distribution in (30.29) is unimportant. For (30.31) use the multidimensional method of moments (Problem 30.6) and the mapping theorem. For (30.32) use the central limit theorem; by (30.28), X_1 has mean 0 and variance $\frac{1}{2}$.

30.10. If $n^{1/2} < m \le n$ and the inequality in (30.33) holds, then $\log\log n^{1/2} < \log\log n - \epsilon(\log\log n)^{1/2}$, which implies $\log\log n < \epsilon^{-2} \log^2 2$. For large n the probability in (30.33) is thus at most $1/\sqrt{n}$.

Section 31

31.1. Consider the argument in Example 31.1. Suppose that F has a nonzero derivative at x, and let I_n be the set of numbers whose base-r expansions agree in the first n places with that of x. The analogue of (31.16) is $P[X \in I_{n+1}]/P[X \in I_n] \to r^{-1}$, and the ratio here is one of $p_0, \ldots, p_r$. If $p_i \ne r^{-1}$ for some i, use the second Borel–Cantelli lemma to show that the ratio is p_i infinitely often except on a set of Lebesgue measure 0. (This last part of the argument is unnecessary if $r = 2$.)

The argument in Example 31.3 needs no essential change. The analogue of (31.17) is

$$F(x) = p_0 + \cdots + p_{i-1} + p_i F(rx - i), \quad \frac{i}{r} \le x \le \frac{i+1}{r}, \quad 0 \le i < r - 1.$$

31.3. (b) Take $f_1 = I_{g^{-1}H_0}$ and $f_2 = F$; $(f_1 f_2)^{-1}\{1\} = H_0$ is not a Lebesgue set.

31.9. Suppose that A is bounded, define μ by $\mu(B) = \lambda(B \cap A)$, and let F be the corresponding distribution function. It suffices to show that $F'(x) = 1$ for x in A, apart from a set of Lebesgue measure 0. Let C_ϵ be the set of x in A for which $F'(x) \le 1 - \epsilon$. From Theorem 31.4(i) deduce that $\lambda(C_\epsilon) = \mu(C_\epsilon) \le (1 - \epsilon)\lambda(C_\epsilon)$ and hence $\lambda(C_\epsilon) = 0$. Thus $F'(x) > 1 - \epsilon$ almost everywhere on A. Obviously, $F'(x) \le 1$.

31.11. Let A be the set of x in the unit interval for which $F'(x) = 0$, take $\alpha = 0$, and define A_n as in the first part of the proof of Theorem 31.4. Choose n so that $\lambda(A_n) \ge 1 - \epsilon$. Split $\{1, 2, \ldots, n\}$ into the set M of k for which $((k-1)/n, k/n]$ meets A_n and the opposite set N. Prove successively that $\sum_{k \in M}[F(k/n) - F((k-1)/n)] \le \epsilon$, $\sum_{k \in N}[F(k/n) - F((k-1)/n)] \ge 1 - \epsilon$, $\sum_{k \in M} 1/n \ge \lambda(A_n) \ge 1 - \epsilon$, $\sum_{k=1}^{n}|f(k/n) - f((k-1)/n)| \ge 2 - 2\epsilon$.

31.15. $\prod_{n=1}^{\infty}(\frac{1}{2} + \frac{1}{2}e^{2it/3^n})$.

31.18. For x fixed, let u_n and v_n be the pair of successive dyadic rationals of order n $(v_n - u_n = 2^{-n})$ for which $u_n < x \le v_n$. Show that

$$\frac{f(v_n) - f(u_n)}{v_n - u_n} = \sum_{k=0}^{n-1}\frac{a_k(v_n) - a_k(u_n)}{v_n - u_n} = \sum_{k=0}^{n-1}a_k^-(x),$$

where a_k^- is the left-hand derivative. Since $a_k^-(x) = \pm 1$ for all x and k, the difference ratio cannot have a finite limit

31.22. Let A be the x-set where (31.35) fails if f is replaced by $f\varphi$; then A has Lebesgue measure 0. Let G be the union of all open sets of μ-measure 0; represent G as a countable disjoint union of open intervals, and let B be G together with any endpoints of zero μ-measure of these intervals. Let D be the set of discontinuity points of F. If $F(x) \notin A$, $x \notin B$, and $x \notin D$, then $F(x - h) < F(x) < F(x + h)$, $F(x \pm h) \to F(x)$, and

$$\frac{1}{F(x+h) - F(x-h)}\int_{F(x-h)}^{F(x+h)}f(\varphi(t))\,dt \to f(\varphi(F(x))).$$

Now $x - \epsilon < \varphi(F(x)) \le x$ follows from $F(x - \epsilon) < F(x)$, and hence $\varphi(F(x)) = x$. If λ is Lebesgue measure restricted to $(0,1)$, then $\mu = \lambda\varphi^{-1}$, and (31.36) follows by change of variable. But (36.36) is easy if $x \in D$, and hence it holds outside $B \cup (D^c \cap F^{-1}A)$. But $\mu(B) = 0$ by construction and $\mu(D^c \cap F^{-1}A) = 0$ by Problem 14.4.

Section 32

32.7. Define μ_n and ν_n as in (32.7), and write $\nu_n = \nu_{ac}^{(n)} + \nu_s^{(n)}$, where $\nu_{ac}^{(n)}$ is absolutely continuous with respect to μ_n and $\nu_s^{(n)}$ is singular with respect to μ_n. Take $\nu_{ac} = \sum_n \nu_{ac}^{(n)}$ and $\nu_s = \sum_n \nu_s^{(n)}$.

Suppose that $\nu_{ac}(E) + \nu_s(E) = \nu_{ac}'(E) + \nu_s'(E)$ for all E in $\mathscr{F}$. Choose an S in $\mathscr{F}$ that supports ν_s and ν_s' and satisfies $\mu(S) = 0$. Then $\nu_{ac}(E) = \nu_{ac}(E \cap S^c) = \nu_{ac}(E \cap S^c) + \nu_s(E \cap S^c) = \nu_{ac}'(E \cap S^c) + \nu_s'(E \cap S^c) = \nu_{ac}'(E \cap S^c) = \nu_{ac}'(E)$. A similar argument shows that $\nu_s(E) = \nu_s'(E)$.

32.8. (a) Show that $\mathscr{B}$ is closed under the formation of countable unions, choose $\mathscr{B}$-sets B_n such that $\mu(B_n) \to \sup_{\mathscr{B}} \mu(B)$ ($< \infty$), and take $B_0 = \bigcup_n B_n$.

(b) The same argument.

(c) Suppose $\mu(D_0) > 0$. The maximality of B_0 implies that $B_0 \cup D_0$ contains an E such that $\mu(E) > 0$ and $\nu(E) < \infty$. Since $B_0 \cap E \subset B_0 \in \mathscr{B}$, $\mu(B_0 \cap E) = 0$ ($\nu(E) < \infty$ rules out $\nu(B_0 \cap E) = \infty$). Therefore, $\mu(D_0 \cap E) > 0$ and $\nu(D_0 \cap E) < \infty$, which contradicts the maximality of C_0.

(d) Take the density to be ∞ on Ω_0^c.

32.9. Define f and ν_s as in (32.8), and let f° and ν_s° be the corresponding function and measure for $\mathscr{F}^\circ$: $\nu(E) = \int_E f^\circ\,d\mu + \nu_s^\circ(E)$ for $E \in \mathscr{F}^\circ$, and there is an $\mathscr{F}^\circ$-set S° such that $\nu_s^\circ(\Omega - S^\circ) = 0$ and $\mu(S^\circ) = 0$. If $E \in \mathscr{F}^\circ$, it follows that $\int_E f^\circ\,d\mu = \int_{E-S^\circ}f^\circ\,d\mu = \int_{E-S^\circ}f^\circ\,d\mu^\circ = \nu^\circ(E - S^\circ) = \nu(E - S^\circ) \ge \int_{E-S^\circ}f\,d\mu = \int_E f\,d\mu$.

It is instructive to consider the extreme case $\mathscr{F}^\circ = \{0, \Omega\}$, in which ν° is absolutely continuous with respect to μ° (provided $\mu(\Omega) > 0$) and hence ν_s° vanishes.

Section 33

33.2. **(a)** To prove independence, check the covariance. Now use Example 33.7.

(b) Use the fact that R and Θ are independent (Example 20.2).

(c) As the single event $[X = Y] = [X - Y = 0] = [\Theta = \pi/4] \cup [\Theta = 5\pi/4]$ has probability 0, the conditional probabilities have no meaning, and strictly speaking there is nothing to resolve. But whether it is natural to regard the degrees of freedom as one or as two depends on whether the $45°$ line through the origin is regarded as an element of the decomposition of the plane into $45°$ lines or whether it is regarded as the union of two elements of the decomposition of the plane into rays from the origin.

Borel's paradox can be explained the same way: The equator is an element of the decomposition of the sphere into lines of constant latitude; the Greenwich meridian is an element of the decomposition of the sphere into great circles with common poles. The decomposition matters, which is to say the σ-field matters.

33.3. **(a)** If the guard says, "1 is to be executed," then the conditional probability that 3 is also to be executed is $1/(1 + p)$. The "paradox" comes from assuming that p must be 1, in which case the conditional probability is indeed $\frac{1}{2}$. But if $p \neq \frac{1}{2}$, then the guard does give prisoner 3 some information.

(b) Here "one" and "other" are undefined, and the problem ignores the possibility that you have been introduced to a girl. Let the sample space be

$$bbo\ \frac{\alpha}{4}, \qquad bgo\ \frac{\beta}{4}, \qquad gbo\ \frac{\gamma}{4}, \qquad ggo\ \frac{\delta}{4},$$

$$bby\ \frac{1-\alpha}{4}, \qquad bgy\ \frac{1-\beta}{4}, \qquad gby\ \frac{1-\gamma}{4}, \qquad ggy\ \frac{1-\delta}{4}.$$

For example, bgo is the event (probability $\beta/4$) that the older child is a boy, the younger is a girl, and the child you have been introduced to is the older; and ggy is the event (probability $(1 - \delta)/4$) that both children are girls and the one you have been introduced to is the younger. Note that the four sex distributions do have probability $\frac{1}{4}$. If the child you have been introduced to is a boy, then the conditional probability that the other child is also a boy is $p = 1/(2 + \beta - \gamma)$. If $\beta = 1$ and $\gamma = 0$ (the parents present a son if they have one), then $p = \frac{1}{3}$. If $\beta = \gamma$ (the parents are indifferent), then $p = \frac{1}{2}$. Any p between $\frac{1}{3}$ and 1 is possible.

This problem shows again that one must keep in mind the entire experiment the sub-σ-field $\mathscr{G}$ represents, not just one of the possible outcomes of the experiment.

33.6. There is no problem, unless the notation gives rise to the illusion that $p(A|x)$ is $P(A \cap [X = x])/P[X = x]$.

33.15. If N is a standard normal variable, then

$$\frac{1}{\sqrt{n}} p_y\left(y + \frac{x}{\sqrt{n}}\right) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} f\left(y + \frac{x}{\sqrt{n}}\right) \bigg/ E\left[f\left(y + \frac{N}{\sqrt{n}}\right)\right].$$

Section 34

34.3. If (X, Y) takes the values $(0, 0)$, $(1, -1)$, and $(1, 1)$ with probability $\frac{1}{3}$ each, then X and Y are dependent but $E[Y \| X] = E[Y] = 0$.

If (X, Y) takes the values $(-1, 1)$, $(0, -2)$, and $(1, 1)$ with probability $\frac{1}{3}$ each, then $E[X] = E[Y] = E[XY] = 0$ and so $E[XY] = E[X]E[Y]$, but $E[Y \| X] = Y \ne 0 = E[Y]$. Of course, this is another example of dependent but uncorrelated random variables.

34.4. First show that $\int f dP_0 = \int_B f dP / P(B)$ and that $P[B \| \mathscr{G}] > 0$ on a set of P_0-measure 1. Let G be the general set in $\mathscr{G}$.

(a) Since

$$\int_G P_0[A \| \mathscr{G}] P[B \| \mathscr{G}] \, dP = \int_G P_0[A \| \mathscr{G}] I_B \, dP = \int_B I_G P_0[A \| \mathscr{G}] \, dP$$

$$= P(B) \int_\Omega I_G P_0[A \| \mathscr{G}] \, dP_0 = P(B) P_0(A \cap G)$$

$$= \int_G P[A \cap B \| \mathscr{G}] \, dP,$$

it follows that

$$P_0[A \| \mathscr{G}] P[B \| \mathscr{G}] = P[A \cap B \| \mathscr{G}]$$

holds on a set of P-measure 1.

(b) If $P_i(A) = P(A | B_i)$, then

$$\int_{G \cap B_i} P_i[A \| \mathscr{G}] \, dP = P(B_i) \int_\Omega I_G P_i[A \| \mathscr{G}] \, dP_i = P(B_i) P_i(A \cap G)$$

$$= \int_{G \cap B_i} P[A \| \mathscr{G} \vee \mathscr{H}] \, dP.$$

Therefore, $\int_C I_{B_i} P_i[A \| \mathscr{G}] \, dP = \int_C I_{B_i} P[A \| \mathscr{G} \vee \mathscr{H}] \, dP$ if $C = G \cap B_i$, and of course this holds for $C = G \cap B_j$ if $j \ne i$. But C's of this form constitute a π-system generating $\mathscr{G} \vee \mathscr{H}$, and hence $I_{B_i} P_i[A \| \mathscr{G}] = I_{B_i} P[A \| \mathscr{G} \vee \mathscr{H}]$ on a set of P-measure 1. Now use the result in part (a).

34.9. All such results can be proved by imitating the proofs for the unconditional case or else by using Theorem 34.5 (for part (c), as generalized in Problem 34.7). For part (a), it must be shown that it is possible to take the integral measurable $\mathscr{G}$.

34.10. (a) If $Y = X - E[X \| \mathscr{G}_1]$, then $X - E[X \| \mathscr{G}_2] = Y - E[Y \| \mathscr{G}_2]$, and $E[(Y - E[Y \| \mathscr{G}_2])^2 \| \mathscr{G}_2] = E[Y^2 \| \mathscr{G}_2] - E^2[Y \| \mathscr{G}_2] \le E[Y^2 \| \mathscr{G}_2]$. Take expected values.

34.11. First prove that

$$P[A_1 \cap A_3 \| \mathscr{G}_2] = E[I_{A_1} P[A_3 \| \mathscr{G}_{12}] \| \mathscr{G}_2].$$

From this and (i) deduce (ii). From

$$E\left[I_{A_1}P[A_3\|\mathscr{G}_2]\|\mathscr{G}_2\right] = P[A_1\|\mathscr{G}_2]P[A_3\|\mathscr{G}_2],$$

(ii), and the preceding equation deduce

$$\int_{A_1\cap A_2} P[A_3\|\mathscr{G}_2]\,dP = \int_{A_1\cap A_2} P[A_3\|\mathscr{G}_{12}]\,dP.$$

The sets $A_1 \cap A_2$ form a π-system generating $\mathscr{G}_{12}$.

34.16. (a) Obviously (34.18) implies (34.17). If (34.17) holds, then clearly (34.18) holds for X simple. For the general X, choose simple X_k such that $\lim_k X_k = X$ and $|X_k| \le |X|$. Note that

$$\left| \int_{A_n} X\,dP - \alpha\!\int X\,dP \right|$$

$$\le \left| \int_{A_n} X_k\,dP - \alpha\!\int X_k\,dP \right| + (1 + |\alpha|)E[|X - X_k|];$$

let $n \to \infty$ and then let $k \to \infty$.

(b) If $\Omega \in \mathscr{P}$, then the class of E satisfying (34.17) is a λ-system, and so by the $\pi-\lambda$ theorem and part (a), (34.18) holds if X is measurable $\sigma(\mathscr{P})$. Since $A_n \in \sigma(\mathscr{P})$, it follows that

$$\int_{A_n} X\,dP = \int_{A_n} E[X\|\sigma(\mathscr{P})]\,dP \to \alpha\!\int E[X\|\sigma(\mathscr{P})]\,dP$$

$$= \alpha\!\int X\,dP.$$

(c) Replace X by $X\,dP_0/dP$ in (34.18).

34.17. (a) The Lindeberg–Lévy theorem.
(b) Chebyshev's inequality.
(c) Theorem 25.4.
(d) Independence of the X_n.
(e) Problem 34.16(b).
(f) Problem 34.16(c).
(g) Part (b) here and the ϵ–δ definition of absolute continuity.
(h) Theorem 25.4 again.

Section 35

35.4. (b) Let S_n be the number of k such that $1 \le k \le n$ and $Y_k = \frac{3}{2}$. Then $X_n = 3^{S_n}/2^n$. Take logarithms and use the strong law of large numbers.

35.9. Let K bound $|X_1|$ and the $|X_n - X_{n-1}|$. Bound $|X_\tau|$ by $K\tau$. Write $\int_{\tau \le k} X_\tau \, dP = \sum_{i=1}^{k} \int_{\tau=i} X_i \, dP = \sum_{i=1}^{k} (\int_{\tau \ge i} X_i \, dP - \int_{\tau \ge i+1} X_i \, dP)$. Transform the last integral by the martingale property and reduce the expression to $E[X_1] - \int_{\tau > k} X_{k+1} \, dP$. Now

$$\left| \int_{\tau > k} X_{k+1} \, dP \right| \le K(k+1) P[\tau > k] \le K(k+1)k^{-1} \int_{\tau > k} \tau \, dP \to 0.$$

35.13. (a) By the result in Problem 32.9, $X_1, X_2, \ldots$ is a supermartingale. Since $E[|X_n|] = E[X_n] \le \nu(\Omega)$, Theorem 35.5 applies.

(b) If $A \in \mathscr{F}_n$, then $\int_A (Y_n + Z_n) \, dP + \sigma_n'(A) = \int_A X_\infty \, dP + \sigma_\infty(A) = \nu(A) = \int_A X_n \, dP + \sigma_n(A)$. Since the Lebesgue decomposition is unique (Problem 32.7), $Y_n + Z_n = X_n$ with probability 1. Since X_n and Y_n converge, so does Z_n. If $A \in \mathscr{F}_k$ and $n \ge k$, then $\int_A Z_n \, dP \le \sigma_\infty(A)$, and by Fatou's lemma, the limit Z satisfies $\int_A Z \, dP \le \sigma_\infty(A)$. This holds for A in $\bigcup_k \mathscr{F}_k$ and hence (monotone class theorem) for A in $\mathscr{F}_\infty$. Choose A so that $P(A) = 1$ and $\sigma_\infty(A) = 0$: $E[Z] = \int_A Z \, dP \le \sigma_\infty(A) = 0$.

It can happen that $\sigma_n(\Omega) = 0$ and $\sigma_\infty(\Omega) = \nu(\Omega) > 0$, in which case σ_n does not converge to σ_∞ and the X_n cannot be integrated to the limit.

35.17. For a very general result, see J. L. Doob: Application of the theory of martingales, *Le Calcul des Probabilités et ses Applications* (Colloques Internationaux du Centre de la Recherche Scientifique, Paris, 1949).

Section 36

36.5. (b) Show by part (a) and Problem 34.18 that f_n is the conditional expected value of f with respect to the σ-field $\mathscr{T}_{n+1}$ generated by the coordinates $x_{n+1}, x_{n+2}, \ldots$. By Theorem 35.9, (36.30) will follow if each set in $\bigcap_n \mathscr{T}_n$ has π-measure either 0 or 1, and here the zero–one law applies.

(c) Show that g_n is the conditional expected value of f with respect to the σ-field generated by the coordinates $x_1, \ldots, x_n$, and apply Theorem 35.6.

36.7. Let $\mathscr{S}$ be the countable set of simple functions $\sum_i \alpha_i I_{A_i}$ for α_i rational and $\{A_i\}$ a finite decomposition of the unit interval into subintervals with rational endpoints. Suppose that the X_t exist, and choose (Theorem 17.1) Y_t in $\mathscr{S}$ so that $E[|X_t - Y_t|] < \frac{1}{4}$. From $E[|X_s - X_t|] = \frac{1}{2}$, conclude that $E[|Y_s - Y_t|] > 0$ for $s \ne t$. But there are only countably many of the Y_t. It does no good to replace Lebesgue measure by some other measure on the unit interval. ✦

Section 37

37.1. If $t_1, \ldots, t_k$ are in increasing order and $t_0 = 0$, then

$$\sum_{i,j} K(t_i, t_j) x_i x_j = \sum_{i,j} x_i x_j \sum_{l=1}^{\min(i,j)} (t_l - t_{l-1})$$

$$= \sum_l (t_l - t_{l-1}) \left(\sum_{i \ge l} x_i \right)^2 \ge 0.$$

37.4 **(a)** Use Problem 36.6(b).

   (b) Let $[W_t: t \geq 0]$ be a Brownian motion on $(\Omega, \mathcal{F}, P_0)$, where $W(\cdot, \omega) \in C$ for every ω. Define $\xi: \Omega \to R^T$ by $Z_t(\xi(\omega)) = W_t(\omega)$. Show that ξ is measurable $\mathcal{F}/\mathcal{R}^T$ and $P = P_0\xi^{-1}$. If $C \subset A \in \mathcal{R}^T$, then $P(A) = P_0(\xi^{-1}A) = P_0(\Omega) = 1$.

37.5. Consider $W(1) = \sum_{k=1}^{n}(W(k/n) - W((k-1)/n))$ for notational convenience. Since

$$n \int_{|W(1/n)| \geq \epsilon} W^2\left(\frac{1}{n}\right) dP = \int_{|W(1)| \geq \epsilon \sqrt{n}} W^2(1) \, dP \to 0,$$

the Lindeberg theorem applies.

37.14. By symmetry,

$$\rho(s, t) = 2P\left[W_s > 0, \inf_{s \leq u \leq t} (W_u - W_s) \leq -W_s\right];$$

W_s and the infimum here are independent because of the Markov property, and so by (20.30) (and symmetry again)

$$\rho(s, t) = 2\int_0^\infty P[\tau_x \leq t - s]\frac{1}{\sqrt{2\pi s}}e^{-x^2/2s} \, dx$$

$$= 2\int_0^\infty \int_0^{t-s} \frac{x}{\sqrt{2\pi}} \frac{1}{u^{3/2}}e^{-x^2/2u} \frac{1}{\sqrt{2\pi s}}e^{-x^2/2s} \, du \, dx.$$

Reverse the integral, use $\int_0^\infty xe^{-x^2 r/2} \, dx = 1/r$, and put $v = (s/(s+u))^{1/2}$:

$$\rho(s, t) = \frac{1}{\pi}\int_0^{t-s} \frac{1}{u+s} \frac{s^{1/2}}{u^{1/2}} \, du$$

$$= \frac{2}{\pi}\int_{\sqrt{s/t}}^1 \frac{dv}{\sqrt{1-v^2}}.$$

Bibliography

Halmos and Saks have been the strongest measure-theoretic and Doob and Feller the strongest probabilistic influences on this book, and the spirit of Kac's small volume has been very important.

Aubrey: *Brief Lives*, John Aubrey; ed., O. L. Dick. Seker and Warburg, London, 1949.

Bahadur: *Some Limit Theorems in Statistics*, R. R. Bahadur. SIAM, Philadelphia, 1971.

Banach: *Théorie des Opérations Linéaires*, S. Banach. Monografje Matematyczne, Warsaw, 1932.

Berger: *Statistical Decision Theory*, 2nd ed., James O. Berger. Springer-Verlag, New York, 1985.

Bhattacharya & Waymire: *Stochastic Processes with Applications*, Rabi N. Bhattacharya and Edward C. Waymire. Wiley, New York, 1990.

Billingsley$_1$: *Convergence of Probability Measures*, Patrick Billingsley. Wiley, New York, 1968.

Billingsley$_2$: *Weak Convergence of Measures: Applications in Probability*, Patrick Billingsley. SIAM, Philadelphia, 1971.

Birkhoff & Mac Lane: *A Survey of Modern Algebra*, 4th ed., Garrett Birkhoff and Saunders Mac Lane. Macmillan, New York, 1977.

Çinlar: *Introduction to Stochastic Processes*, Erhan Çinlar. Prentice-Hall, Englewood Cliffs, New Jersey, 1975.

Cramér: *Mathematical Methods of Statistics*, Harald Cramér. Princeton University Press, Princeton, New Jersey, 1946.

Doob: *Stochastic Processes*, J. L. Doob. Wiley, New York, 1953.

Dubins & Savage: *How to Gamble If You Must*, Lester E. Dubins and Leonard J. Savage. McGraw-Hill, New York, 1965.

DUDLEY: *Real Analysis and Probability*, Richard M. Dudley. Wadsworth and Brooks, Pacific Grove, California, 1989.

DYNKIN & YUSHKEVICH: *Markov Processes*, English ed., Evgenii B. Dynkin and Aleksandr A. Yushkevich. Plenum Press, New York, 1969.

FELLER: *An Introduction to Probability Theory and Its Applications*, Vol. I. 3rd ed., Vol. II, 2nd ed., William Feller. Wiley, New York, 1968, 1971.

GALAMBOS: *The Asymptotic Theory of Extreme Order Statistics*, Janos Galambos. Wiley, New York, 1978.

GELBAUM & OLMSTED: *Counterexamples in Analysis*, Bernard R. Gelbaum and John M. Olmsted. Holden-Day, San Francisco, 1964.

GNEDENKO & KOLMOGOROV: *Limit Distributions for Sums of Independent Random Variables*, English ed., B. V. Gnedenko and A. N. Kolmogorov. Addison-Wesley, Reading, Massachusetts, 1954.

HALMOS₁: *Measure Theory*, Paul R. Halmos. Van Nostrand, New York, 1950.

HALMOS₂: *Naive Set Theory*, Paul R. Halmos. Van Nostrand, Princeton, 1960.

HARDY: *A Course of Pure Mathematics*, 9th ed., G. H. Hardy. Macmillan, New York, 1946.

HARDY & WRIGHT: *An Introduction to the Theory of Numbers*, 4th ed., G. H. Hardy and E. M. Wright. Clarendon, Oxford, 1959.

HAUSDORFF: *Set Theory*, 2nd English ed., Felix Hausdorff. Chelsea, New York, 1962.

JECH: *Set Theory*, Thomas Jech. Academic Press, New York, 1978.

KAC: *Statistical Independence in Probability, Analysis and Number Theory*, Carus Math. Monogr. 12, Marc Kac. Wiley, New York, 1959.

KAHANE: *Some Random Series of Functions*, Jean-Pierre Kahane. Heath, Lexington, Massachusetts, 1968.

KAPLANSKY: *Set Theory and Metric Spaces*, Irving Kaplansky. Chelsea, New York, 1972.

KARATZES & SHREVE: *Brownian Motion and Stochastic Calculus*, Ioannis Karatzis and Steven E. Shreve. Springer-Verlag, New York, 1988.

KARLIN & TAYLOR: *A First Course in Stochastic Processes*, 2nd ed., *A Second Course in Stochastic Processes*, Samuel Karlin and Howard M. Taylor. Academic Press, New York, 1975 and 1981.

KOLMOGOROV: *Grundbegriffe der Wahrscheinlichkeitsrechnung*, Erg. Math., Vol. 2, No. 3, A. N. Kolmogorov. Springer-Verlag, Berlin, 1933.

LÉVY: *Théorie de l'Addition des Variables Aléatoires*, Paul Lévy. Gauthier-Villars, Paris, 1937.

PROTTER: *Stochastic Integration and Differential Equations*, Philip Protter. Springer-Verlag, 1990.

RIESZ & SZ.-NAGY: *Functional Analysis*, English ed., Frigyes Riesz and Bela Sz.-Nagy. Unger, New York, 1955.

ROCKETT & SZÜSZ: *Continued Fractions*, Andrew M. Rockett and Peter Szüsz. World Scientific, Singapore, 1992.

ROYDEN: *Real Analysis*, 2nd ed., H. I. Royden. Macmillan, New York, 1968.

RUDIN$_1$: *Principles of Mathematical Analysis*, 3rd ed., Walter Rudin. McGraw-Hill, New York, 1976.

RUDIN$_2$: *Real and Complex Analysis*, 2nd ed., Walter Rudin. McGraw-Hill, New York, 1974.

SAKS: *Theory of the Integral*, 2nd rev. ed., Stanislaw Saks. Hafner, New York, 1937.

SPIVAK: *Calculus on Manifolds*, Michael Spivak. W. A. Benjamin, New York, 1965.

WAGON: *The Banach–Tarsky Paradox*, Stan Wagon. Cambridge University Press, 1985.

List of Symbols

Index

Here An refers to paragraph n of the Appendix (p. 536): $u.v$ refers to Problem v in Section u, or else to a note on it (Notes on the Problems, p. 552); the other references are to pages. Greek letters are alphabetized by their Roman equivalents (m for μ, and so on). Names in the bibliography are not indexed separately.

PROBABILITY AND MEASURE
Third Edition

Now in its new third edition, *Probability and Measure* offers advanced students, scientists, and engineers an integrated introduction to measure theory and probability. Retaining the unique approach of the previous editions, this text interweaves material on probability and measure, so that probability problems generate an interest in measure theory and measure theory is then developed and applied to probability. *Probability and Measure* provides thorough coverage of probability, measure, integration, random variables and expected values, convergence of distributions, derivatives and conditional probability, and stochastic processes. The *Third Edition* features an improved treatment of Brownian motion and the replacement of queuing theory with ergodic theory.

Like the previous editions, this new edition will be well received by students of mathematics, statistics, economics, and a wide variety of disciplines that require a solid understanding of probability theory.

PATRICK BILLINGSLEY is Professor of Statistics and Mathematics at the University of Chicago. He is the coauthor (with Watson et al.) of *Statistics for Management and Economics*; (with D. L. Huntsberger) of *Elements of Statistical Inference*; and the author of *Convergence of Probability Measures* (Wiley-Interscience), among other works. Dr. Billingsley has also edited the *Annals of Probability* for the Institute of Mathematical Statistics. He received his PhD in mathematics from Princeton University.