Introduction to Clinical Trials

Lecture 5: Power and Sample Size Calculations

Giorgos Bakoyannis, PhD
Associate Professor
Department of Biostatistics and Health Data Science
Indiana University Indianapolis

Notes by Prof. Constantin T. Yiannoutsos, City University of New York

Section 1

Introduction

Terminology

The following concepts will be referred to repeatedly in this lecture:

Power:	1-β
β level:	Type II error probability
α level:	Type I error probability
Likelihood ratio:	Relative strength of evidence
Sample size:	Number of experimental subjects
Effect size:	Treatment difference expressed as the number of standard deviations
Number of events:	Number of subjects with a specific outcome
Study duration:	Time from beginning of the trial to end of follow-up
Percent censoring:	Proportion of participants without an event by the end of the study
Allocation ratio:	Ratio of sample size in the treatment groups
Accrual rate:	New subjects entered per unit of time
Loss to follow-up rate:	Rate at which study participants are lost before outcomes are observed
Follow-up period:	Interval from end of accrual to end of study
Δ:	Smallest treatment effect of interest based on clinical considerations

Power

Power is the chance that a true difference will be detected by the study. There are a number of conceptual difficulties with this:

- Power is defined hypothetically (the treatment effect is actually present) as opposed to the null hypothesis of no effect or treatment difference
- Power is related to the experiment-wide variability
- Power cannot be separated by the sample size and the treatment effect. Thus, statements like “this study produced 90% power” are erroneous and possibly misleading. In fact *any* study can be made to generate any level of power (just assume a larger effect)

Section 2

Early developmental studies

Early developmental trials
Translational studies

In translational trials, the sample size n is such that it would ensure that the absolute error lies below some threshold d with high probability. In other words,

$$P(|\bar{X}_n - \mu| \leq d) \geq 1 - \alpha$$

which is equivalent to

$$P\left(-\frac{d}{\sigma/\sqrt{n}} \leq \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq \frac{d}{\sigma/\sqrt{n}}\right) \geq 1 - \alpha.$$

Early developmental trials
Translational studies

By the central limit theorem we have that (for sufficiently large sample size n)

$$\begin{aligned} P\left(-\frac{d}{\sigma/\sqrt{n}} \leq \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq \frac{d}{\sigma/\sqrt{n}}\right) &\approx \Phi\left(\frac{d}{\sigma/\sqrt{n}}\right) - \Phi\left(-\frac{d}{\sigma/\sqrt{n}}\right) \\ &= 2 \times \Phi\left(\frac{d}{\sigma/\sqrt{n}}\right) - 1 \end{aligned}$$

where $\Phi(z) = P(Z \leq z)$ with $Z \sim N(0,1)$. Thus,

$$2 \times \Phi\left(\frac{d}{\sigma/\sqrt{n}}\right) - 1 \geq 1 - \alpha$$

$$\Rightarrow \Phi\left(\frac{d}{\sigma/\sqrt{n}}\right) \geq 1 - \alpha/2$$

$$\Rightarrow \frac{d}{\sigma/\sqrt{n}} \geq z_{1-\alpha/2}$$

Early developmental trials
Translational studies (continued)

Thus, the required sample size n is

$$n \geq \left(z_{1-\alpha/2}\frac{\sigma}{d}\right)^2$$

The above can be expressed in terms of *effect size*. That is, we may want to calculate the sample size required to ensure (at the $\alpha = 0.05$ say), that the error $d = 0.5\sigma$ (or, equivalently, $d/\sigma = 0.5$). The above formula is then

$$\begin{aligned} n &= \left(z_{1-\alpha/2}\frac{\sigma}{d}\right)^2 \\ &= (1.96/0.5)^2 \approx 16 \end{aligned}$$

Early developmental trials
Testing a single mean

When the mean is tested then the one-sided null and alternative hypotheses are as follows:

$$H_0 : \mu \leq \mu_0 \text{ or } H_0 : \mu \geq \mu_0$$

versus, respectively,

$$H_A : \mu > \mu_0 \text{ or } H_A : \mu < \mu_0$$

or the two-sided null and alternative

$$H_0 : \mu = \mu_0$$

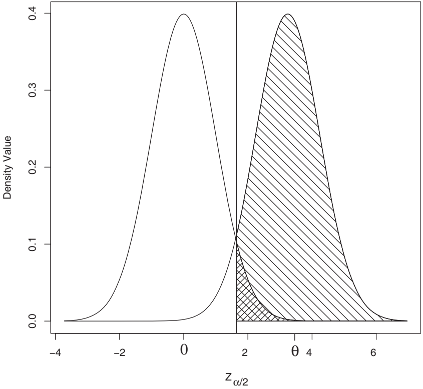
versus

$$H_A : \mu \neq \mu_0$$

Early developmental trials
Testing a single mean (continued)

Testing the null against the one-sided alternative $H_A : \mu = \mu_1 > \mu_0$, for a pre-specified value of μ_1 the test is based on the sample distribution of the mean $\bar{X}_n$ which, under H_0 is $N(\mu_0, \sigma^2/n)$ and under H_A $N(\mu_1, \sigma^2/n)$.

This is shown pictorially in the Figure below:



Early developmental trials
Sample size for testing a single mean

In light of the central limit theorem, the null hypothesis is rejected at the level α if

$$\frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} \geq z_{1-\alpha}$$

If the true mean is μ_1 (i.e., under H_A) then

$$\begin{aligned} P\left(\frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} \geq z_{1-\alpha}\right) &= P\left(\bar{X}_n \geq z_{1-\alpha}\frac{\sigma}{\sqrt{n}} + \mu_0\right) \\ &= P\left(\frac{\bar{X}_n - \mu_1}{\sigma/\sqrt{n}} \geq z_{1-\alpha} - \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}}\right) \\ &\approx 1 - \Phi\left(z_{1-\alpha} - \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}}\right) \end{aligned}$$

for sufficiently large sample size n (by the central limit theorem).

Early developmental trials
Sample size for testing a single mean (continued)

Letting $1 - \beta$ be the desired power we have that

$$1 - \Phi\left(z_{1-\alpha} - \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}}\right) \geq 1 - \beta$$

$$\Rightarrow \Phi\left(z_{1-\alpha} - \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}}\right) \leq \beta$$

$$\Rightarrow z_{1-\alpha} - \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}} \leq z_\beta$$

$$\Rightarrow n \geq \frac{\sigma^2(z_\alpha + z_\beta)^2}{(\mu_1 - \mu_0)^2}$$

Note that we can express the equation above in terms of effect size $f = (\mu_1 - \mu_0)/\sigma$, i.e., $n = (z_\alpha + z_\beta)^2/f^2$.

Early developmental trials
Sample size for testing a single mean (continued)

As an example, consider the sample size required for a test of the hypothesis $H_0 : \mu \leq 3$ versus the alternative $H_A : \mu > \mu_0 = 4$ (this is completely contrived example). If the α and β levels are, respectively, 5% and 10% (or the power is 90%) and the standard deviation $\sigma = 2$, the required sample size will be

$$n = \frac{2^2(-1.645 - 1.282)^2}{(4-3)^2} \approx 34.27$$

We choose the smallest integer that is greater than or equal to 34.27, so that $n = 35$. The same results would be generated by considering this difference in treatment means as an effect size $f = 0.5$.

Finally, in the case of a two-sided alternative hypothesis, the sample size would be $n = \frac{(-1.96-1.282)^2}{0.5^2} \approx 34$.

Early developmental trials
Sample size for testing a single proportion

The above results can be modified by substituting $\sigma = \sqrt{p(1-p)}$ in the previous formula when testing for a single proportion (of toxicity or response) p. In this case, the null hypothesis is $H_0 : p \leq p_o$ versus the one-sided alternative $H_A : p = p_1 > p_0$. The required sample size is

$$n = \frac{p_0(1-p_0)(z_\alpha + z_\beta)^2}{(p_1 - p_0)^2}$$

Thus, testing the null and alternative hypotheses $H_0 : p \leq 0.3$ and $H_A : p = 0.4$ respectively at the $\alpha = 0.05$ and $\beta = 0.10$ we have

$$n = \frac{0.3(1-0.3)(-1.645 - 1.282)^2}{(0.4-0.3)^2} \approx 180$$

Testing a single proportion
Single-stage design

Consider the following situation:

In a Phase II non-comparative study (i.e, a small study of one treatment that takes a first "stab" at efficacy assessment), we would like to know whether the true response rate is *at least* as high as 15% (the current standard).

Above that rate, the new therapy would be interesting and worth pursuing further, while, below this rate, we would discontinue development of the experimental therapy.

To perform power and sample size calculations we will have to specify an alternative rate $p_1 > p_0$. We set for this example, $p_1 = 0.40$.

Testing a single proportion
Statistical construction

The null hypothesis to be tested is

$$H_0 : p \leq p_0 = 0.15$$

$$H_A : p > p_1 = 0.40$$

versus the alternative

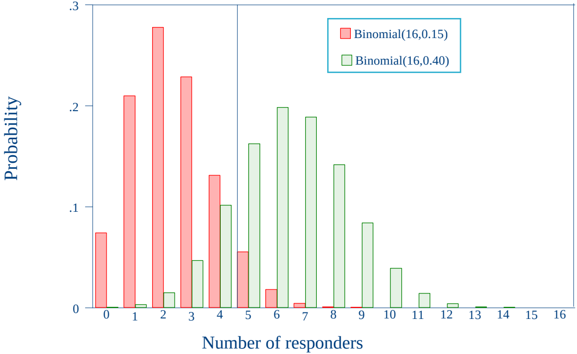
Let's say that we would like to maintain $\alpha \leq 0.1$ and the power $1 - \beta \approx 0.80$. We will thus create a one-sided 90% confidence interval with a lower bound and see whether this lower bound excludes (lies above) $p_0 = 0.15$.

The power is the chance that the confidence interval will lie above p_0 if the true response is p_1.

Monitoring toxicity is done in an identical manner by reversing the roles of p_0 and p_1 and H_0 and H_A above.

Testing a single proportion
Graphical representation of the problem

The situation is given graphically by the following figure:



Of course, unlike the continuous case, the above figure would not have been possible if we had not already determined the size n (see below for how this was done).

Testing a single proportion
Exact binomial confidence intervals

To determine the power at a specific sample size (or vice versa) we use the exact binomial distribution (e.g., Korn, *Biometrics*, 1986).

We work similarly to the one-sample continuous-data case by trying to identify a cutoff point x such that $P(X \leq r|H_0) \geq 1 - \alpha$ and $P(X \leq r|H_A) \leq \beta$. That is,

$$\sum_{k=0}^{r} \left(\begin{array}{c} n \\ k \end{array} \right) p_0^k (1 - p_0)^{n-k} \geq 1 - \alpha$$

and

$$\sum_{k=0}^{r} \left(\begin{array}{c} n \\ k \end{array} \right) p_1^k (1 - p_1)^{n-k} \leq \beta$$

Testing a single proportion
Example

After some experimentation we end up with $n = 16$. In that case, the null hypothesis will be excluded if the number of responding patients is $X \geq 5$.

The alpha level of the test is the chance that $X \geq 5$ under the null hypothesis,

$$P(X \geq 5|H_0) = P(X \geq 5|n = 16, p = 0.15) = 0.0731 = \alpha$$

The power (or the chance that the CI will lie above p_0) is the same probability under the alternative hypothesis i.e.,

$$P(X \geq 5|H_A) = P(X \geq 5|n = 16, p = 0.40) = 0.8334 = 1 - \beta$$

Testing a single proportion
Example

The null and alternative distributions $B(16, 0.15)$ and $B(16, 0.40)$:

k	$P(X = r\|H_0)$	$P(X \leq r\|H_0)$	$P(X = r\|H_A)$	$P(X \leq r\|H_A)$
0	.0743	.0742511	.0002821	.0002821
1	.2097	.2839012	.0030092	.0032913
2	.2775	.5613793	.0150459	.0183372
3	.2285	.7898907	.0468095	.0651467
4	.1311	$\boxed{\mathbf{.9209}}$	.1014206	$\boxed{\mathbf{.1666}}$
5	.0555	.9764556	.162273	.3288404
6	.0180	.9944137	.1983337	.5271741
7	.0045	.998941	.1888892	.7160634
8	.0009	.9998398	.1416669	.8577303
9	.0001	.9999807	.0839508	.941681
10	.0000	.9999982	.039177	.9808581
11	.0000	.9999999	.0142462	.9951043
12	.0000	1	.0039573	.9990615
13	.0000	1	.0008117	.9998733
14	.0000	1	.000116	.9999893
15	.0000	1	.0000103	.9999996
16	.0000	1	.0000000	1

Thus, the alpha level is $\alpha = 1 - 0.9209 = 0.0791$ and the power is $1 - \beta = 0.8334$.

Single-stage designs
The clinfun R package

There is an R package that performs much of these calculations. It is called clinfun and performs many clinical trial design calculations. The general invocation of function ph2single in this package is

```
ph2single(p0, p1, a, b, nsoln)
```

where `nsoln` provides a number of designs which would satisfy the α level and power requirement (although, usually the one with the smallest sample size is chosen).

Single-stage designs
The clinfun R package

In our example, this is as follows:

```
> ph2single(.15, .4, .1, .2, nsoln=5)
   n r Type I error Type II error
1 16 4   0.07905130    0.16656738
2 17 4   0.09871000    0.12599913
3 19 5   0.05369611    0.16292248
4 20 5   0.06730797    0.12559897
5 21 5   0.08273475    0.09574016
```

for 5 designs. The first one is the design produced by the other software we considered earlier.

Testing a single proportion
Two-stage designs

Consider what would be required in order to be able to insert an interim analysis (a first "stage") in the study monitoring. The components of a two-stage design are the following:

- Hypotheses to be tested
  - $H_0 : p \leq p_0$
  - $H_A : p \geq p_A$
- Type I (α), type II (β) errors and power $(1 - \beta)$
- Sample size (n) and total number of responses (r)
  - Stage I: Sample size (n_1) and number of responses (r_1)
  - Stage II: Sample size (n_2) and number of responses (r_2)

Two-stage designs attempt to control the alpha level and power.

Testing a single proportion
Simon's two-stage design

Simon (*Cont Clin Trials*, 1985) proposed the following design:

- **First stage**
  The study is stopped after the first stage for insufficient efficacy if r_1 or less responses out of n_1 total subjects are observed. The probability of early termination under rate p is

$$PET(p) = P(X \leq r_1 | n, p)$$

- **Second stage**
  The study is continued to the second stage if more than r_1 out of n_1 subjects respond during the first stage.
- The study is considered successful (H_0 is rejected) if more than $r = r_1 + r_2$ out of N subjects respond by the end of the second stage.

Simon's two-stage design
Expected versus maximum sample size

In the two-stage design the maximum sample size is random. The *expected* sample size (also known as *average sample number* or ASN) under rate p is given by the following formula:

$$\boxed{\begin{array}{rcl} ASN(p) & = & n_1 + n_2 \times P(k > r_1 | n_1, p) \\ & = & n_1 + n_2 \times (1 - PET(p)) \end{array}}$$

that is, the average sample size equals the number of subjects to be enrolled in the first stage, times the number of subjects enrolled in the second stage probability of continuing to the second stage.

Simon's design minimizes $ASN(p_o)$ (i.e., under the null hypothesis)

The *minimax* design minimizes the maximum sample size n.

Simon's two-stage design
Implementation of the previous example

For example, consider the two-stage design with $n_1 = 9$, $r_1 = 1$, $n = 16$ and $r = 4$. Then under the null hypothesis $p = 0.15$ we have

n	$P(X = k \mid n = 9, p = 0.15)$	$P(X \leq k \mid n = 9, p = 0.15)$
0	0.23162	0.23162
1	0.36786	0.59948
2	0.25967	0.85915
3	0.10692	0.96607
4	0.02830	0.99437
5	0.00499	0.99937
6	0.00059	0.99995
7	0.00004	1.00000
8	0.00000	1.00000
9	0.00000	1.00000

With $r_1 = 1$ and $n_2 = 7$, $r_2 = 3$ (so that $r = 4$) we have

Probability of response $p = 0.15$			
Stage I ($n_1 = 9$) responses	Stage II ($n_2 = 7$) responses	Probability	Cum. prob.
0		0.2316	0.2316
1		0.3679	0.5995
2	0	0.0832	0.6827
	1	0.1028	0.7856
	2	0.0544	0.8400
3	0	0.0343	0.8743
	1	0.0423	0.9166
4	0	0.0091	0.9257

Simon's two-stage design
Attained size of the test

The probability of not rejecting the null hypothesis $H_0 : p \leq 0.15$ when this is true is $1 - \alpha = 0.9257)$. The cumulative probability 0.9257 above is the total probability associated with all scenarios of non-rejection of H_0. These are:

- **First stage**
  The number of responses is $k \leq r_1 = 1$, i.e., $k = 0$, or 1 (this would result in stopping the trial).
- **Second stage**
  In order to proceed to the second stage, $k > 1$. In order *not* to reject the null hypothesis, $k \leq r = 4$, i.e., $k = 2, 3, 4$. The probability is given by summing the binomial probabilities of the compatible scenarios.

Thus, the attained size of the test is $\alpha = 1 - 0.9257 = 0.0743$.

Simon's two-stage design (cont'd)

To estimate power we run the same routine with $p = p_A = 0.40$. The results are as follows:

n	$P(X = k \mid n = 9, p = 0.40)$	$P(X \leq k \mid n = 9, p = 0.40)$
0	0.01008	0.01008
1	0.06047	0.07054
2	0.16124	0.23179
3	0.25082	0.48261
4	0.25082	0.73343
5	0.16722	0.90065
6	0.07432	0.97497
7	0.02123	0.99620
8	0.00354	0.99974
9	0.00026	1.00000

Simon's two-stage design (cont'd)

With $r_1 = 1$ and $n_2 = 7$, $r_2 = 3$ (so that $r = 4$) we have

Probability of response $p = 0.40$

Stage I ($n_1 = 9$) responses	Stage II ($n_2 = 7$) respnses	Probability	Cum. prob.
0		0.0101	0.0101
1		0.0605	0.0705
2	0	0.0045	0.0751
	1	0.0211	0.0961
	2	0.0421	0.1383
3	0	0.0070	0.1453
	1	0.0328	0.1780
4	0	0.0070	0.1851

Simon's two-stage design
Power

The previous output is interpreted as follows:

The probability of not rejecting the null hypothesis $H_0 : p \leq p_0 = 0.15$ when this is false (i.e., the Type II of this test) is $\beta = 0.1851$.

The cumulative probability 0.1851 is given in a manner similar to the calculation of α above by summing the binomial probabilities of the compatible scenarios, but with $p = p_A$ in this case.

The projected power of this study is $1 - \beta = 0.8149$.

Simon's two-stage design
Average sample size

From the output above we can calculate that

- **Under the null hypothesis**

$$\begin{aligned} ASN(p_o) &= n_1 + n_2 \times (1 - B(r_1; n_1, p_o)) \\ &= 9 + 7 \times (1 - 0.59948) = 11.803 \end{aligned}$$

- **Under the alternative hypothesis**

$$\begin{aligned} ASN(p_A) &= n_1 + n_2 \times (1 - B(r_1; n_1, p_A)) \\ &= 9 + 7 \times (1 - 0.07054) = 15.506 \end{aligned}$$

The fact that the expected sample size of the two-stage design (under the null hypothesis) is significantly lower than the sample size of the comparable one-stage design is a critical advantage of this design.

Simon's two-stage design
The R package clinfun

To carry out a Simon or minimax two-stage design, we use the function ph2simon within the R package clinfun. The invocation of this function is

ph2simon(p0, p1, a, b, nmax)

where nmax specifies the maximum n and is 100 by default unless you otherwise specify. In our previous example, the results are as follows:

```
 ph2simon(.15, .4, .1, .2)

 Simon 2-stage Phase II design

Unacceptable response rate:  0.15
Desirable response rate:  0.4
Error rates: alpha =  0.1 ; beta =  0.2

        r1 n1 r  n EN(p0) PET(p0)
Optimal  1  7 4 18  10.12  0.7166
Minimax  1  9 4 16  11.80  0.5995
```

Section 3

Comparative studies

Comparative studies
Testing the difference of two means

In the two-sample case, the null hypothesis is (usually) $H_0 : \mu_1 = \mu_2$. This is equivalent to difference of the two means $\hat{\Delta} = \bar{X}_1 - \bar{X}_2$ under some *a-priori* assumptions.

The distribution of the sample difference of two means, assuming two equal-size $n_1 = n_2 = n$ (say) independent samples and known and equal variances ($\sigma_1^2 = \sigma_2^2 = \sigma^2$) is $\hat{\Delta}_n \dot{\sim} N(\Delta, \sigma_\Delta^2)$, where $\sigma_\Delta^2 = \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$ and $\sigma_\Delta^2 = \left(\frac{2\sigma^2}{n} \right)$ if $n_1 = n_2 = n$.

In other words, the approach is the same as in the single-mean case, with the recognition that the variance is roughly double that of the one-sample case (to acknowledge the estimation of both μ_1 and μ_2).

Comparative studies
Sample size calculations

To calculate the sample size for each group n' we can use the previous one-sample formula, with the appropriate estimate of the variance of course. That is, each group will be comprised of individuals from each population,

$$n' \;=\; \left[\frac{(z_\alpha + z_\beta)}{\Delta}\sigma_\Delta\right]^2 = \left[\frac{(z_\alpha + z_\beta)}{\Delta}\sigma\sqrt{2}\right]^2 = 2\left[\frac{(z_\alpha + z_\beta)}{\Delta}\sigma\right]^2 = 2n$$

where n is the size of the identically defined one-sample case. That is, the sample size in each group in the two-sample case will be roughly double that of the one-sample case.

Comparative studies
Sample size calculations: effect size

We can also express the above formula in terms of the effect size $f = \Delta/\sigma$. In this case, the sample size for each group will be

$$n' = 2\left(\frac{z_\alpha + z_\beta}{f}\right)^2$$

For example, if $f = 0.25$, $\alpha = 0.05$ and $\beta = 0.1$ the required sample size will be

$$n' = 2\left[\frac{(1.645 + 1.282)}{0.25}\right]^2 \approx 275$$

per group for one-sided alternative hypotheses and

$$n' = 2\left[\frac{(1.96 + 1.282)}{0.25}\right]^2 \approx 337$$

per group for two-sided alternatives (see Piantadosi, 2005, pp.280).

Comparing means in a cluster randomization studies

In cluster randomization studies, the sampling unit is the cluster, rather than the individual. Power and sample size calculations in this context need to take into account the cluster randomization design.

To compare two means within a cluster randomization design, we must take into account both the within-cluster variability *and* the between-cluster variability to come up with the estimate of the total variability σ_c^2 to compare mean μ_1 to μ_2 in this design.

Loss of efficiency in the cluster-randomized design

In randomized designs (where the individual observation is the primary sampling unit) the only source of variability is the variability between the units (i.e., σ^2_Δ). Since, in cluster randomized studies, we have an extra source of variability, the variability *between* the clusters.

Thus, all else being equal, in cluster-randomized studies the sample size will be larger than in randomized studies (meaning that there is a, potentially significant, loss of *efficiency* in cluster-randomized studies).

The design effect

The loss of precision (i.e., increase in variance) due to the cluster-randomized trial design is quantified by the *design effect*:

$$Deff = 1 + (m-1)\rho$$

where m is the average number of individuals in each cluster and ρ the intracluster correlation.

The *effective sample size* n_{eff} is the sample size based on simple (non-cluster) randomization that whould achieve the same variance σ^2_Δ of the estimator $\hat{\Delta}_n$ to that in the cluster-randomized clinical trial. The effective sample size is

$$n_{eff} = \frac{n}{Deff}$$

n_{eff} can be used for power calculations in cluster-randomized trials using formulas and software for power analysis based on simple randomization.

Example

The sample size for a number of choices of m and n is given in the following table and is compared to the sample size of the simple randomized study:

No. of patients per practice m	Standard Deviation	No. of practices	No. of patients	Design effect
10	0.364	558	5,580	1.04
25	0.236	234	5,850	1.09
50	0.173	126	6,300	1.17
100	0.132	74	7,400	1.38
500	0.085	32	16,000	2.98
No. needed with individual randomization			5,364	1.00

The conclusion from these results is that the higher the m and the lower the n, the more significant the inefficiency of the design. Conversely, the closer the study comes to an individual randomized design (i.e., m gets close to the total sample size) the less the inefficiency.

Comparative studies
Testing for the difference in two proportions

In the two-sample case, the null hypothesis is (usually) $H_0 : \pi_1 = \pi_2$. This is equivalent to $H_0 : \delta = \pi_1 - \pi_2 = 0$.

Estimation of the difference of the true population proportions $\delta = \pi_1 - \pi_2$ is carried out by using the difference of the two sample proportions $\hat{\delta} = p_1 - p_2$ (where $p_1 = x_1/n_1$ and $p_2 = x_2/n_2$, i.e., the number of successes out of n_1 and n_2 participants in the two groups respectively).

Testing for the difference of two proportions (cont'd)

Under some suitable assumptions, the distribution of the difference of the two sample proportions is $\hat{\delta}_n \dot{\sim} N(0, \sigma_\delta^2)$ under the null hypothesis and $N(\delta, \sigma_\delta^2)$ under the alternative hypothesis, where the variance is $\sigma_\delta^2 = \pi_1(1-\pi_1)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$ under the null hypothesis (with $\sigma_\delta^2 = \frac{2\pi_1(1-\pi_1)}{n}$ when $n_1 = n_2 = n$), and $\sigma_\delta^2 = \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}$ under the alternative (and $\sigma_\delta^2 = \frac{\pi_1(1-\pi_1)+\pi_2(1-\pi_2)}{n}$ when $n_1 = n_2 = n$).

Testing for the difference in two proportions
Sample size calculations

With the exception of the fact that the variance is not the same under the null and alternative hypothesis (and, in fact, that the variance is a function of the unknown quantities π_1 and π_2), the approach is the same as in all previous illustrations.

To calculate the sample size for each group n (unequal sample sizes are handled fairly easily) we use a similar formula to the single-mean case, where, under the null hypothesis, $\pi_1 = \pi_2 = \pi$.

Testing for the difference in two proportions

The following is the formula for the sample size per group in the two-proportion case:

$$n = \frac{\left\{ z_{1-\alpha}\sqrt{2\pi(1-\pi)} - z_{\beta}\sqrt{\pi_1(1-\pi_1)+\pi_2(1-\pi_2)} \right\}^2}{(\pi_2-\pi_1)^2}$$

where, under the null hypothesis, $\pi_1 = \pi_2 = \pi$.

Testing for the difference in two proportions
Example

For example, if $\pi_1 = 0.3$, $\pi_2 = 0.4$, $\alpha = 0.05$ and $\beta = 0.1$ (power=90%), $p = 1/2(0.3 + 0.4) = 0.35$ and the required sample size will be

$$n = \frac{\left\{1.645\sqrt{2(0.35)(0.65)} + 1.282\sqrt{(0.3)(0.7) + 0.4(0.6)}\right\}^2}{(0.4 - 0.3)^2} \approx 388$$

per group for a one-sided alternative hypothesis and

$$n = \frac{\left\{1.96\sqrt{2(0.35)(0.65)} + 1.282\sqrt{(0.3)(0.7) + (0.4)(0.6)}\right\}^2}{(0.4 - 0.3)^2} \approx 477$$

per group for two-sided alternatives.

Testing for the difference in two proportions Equality of means versus lack of association in a 2×2 table

The discussion here is a direct consequence of the 2×2 table setup, which is given below. Considering the "outcome" in the table as success (e.g., death, remission, toxicity, etc.) versus "failure") in the two groups, the table is set up as follows:

	Group		
Outcome	Group 1	Group 2	Total
Success	x_1	x_2	$x_1 + x_2$
Failure	$n - x_1$	$n - x_2$	$n - (x_1 + x_2)$
Total	n_1	n_2	$n = n_1 + n_2$

Then, the hypothesis of no difference between the two proportions (of "Success") is the same as the lack of *association* the outcome and membership in Group 1 or 2.

Testing for the difference in two proportions
Using the Fisher's exact test

One way to address the case of lack of association is to use the, so-called, Fisher's exact test. According to this setup, the margins of the table above (i.e., the row and column totals) are considered fixed. Then the cell counts can be thought of as random draws of n_1 total colored balls from an urn, of which $x_1 + x_2$ have a certain color. x_1 has a *hypergeometric distribution*

$$P(X = x_1) = \frac{\begin{pmatrix} x_1 + x_2 \\ x_1 \end{pmatrix} \begin{pmatrix} n - (x_1 + x_2) \\ n - x_1 \end{pmatrix}}{\begin{pmatrix} n \\ n_1 \end{pmatrix}}$$

We can use the Fisher's exact test to calculate sample sizes in the previous example (this would be preferred, especially, in cases where the sample sizes are small and thus the normal approximation might not be accurate).

Testing for the difference in two proportions Example using the Fisher's exact test in the R package clinfun

The R package clinfun has an array of different programs that use the 2×2 table setup. The function which corresponds to the Fisher's exact test is fe.ssize and is invoked as follows (shown are the default entries):

fe.ssize(p1, p2, alpha=0.05,power=0.8,r=1,npm=5,mmax=1000)

where r is the allocation ratio, npm is a range of $n\pm$npm where the sample calculation search will be conducted and mmax is the maximum group size.

Testing for the difference in two proportions Using the fe.ssize in the previous example

Using the function fe.ssize to calculate the sample size in the previous example we get

```
fe.ssize(.3, .4, alpha=0.1,power=0.9,r=1,npm=5,mmax=1000)
             Group 1 Group 2 Exact Power
CPS              408     408   0.9001173
Fisher Exact     408     408   0.9001173
```

for a one-sided alternative hypothesis (note that the routine will always give the two-sided alternative sample size so the alpha level must be doubled to get the one-sided sample size) and

```
fe.ssize(.3, .4, alpha=0.05,power=0.9,r=1,npm=5,mmax=1000)
             Group 1 Group 2 Exact Power
CPS              496     496   0.9003782
Fisher Exact     496     496   0.9003782
```

for a two-sided alternative. Along with the sample size corresponding to the Fisher's exact test, we also get the Casagrande, Pike and Smith approximation (*Biometrics*, 1978).

The concept of statistical information

The concept of statistical information is central to frequentist analysis. In general, the information about a parameter δ is

$$I \propto \left[\mathsf{Var}(\hat{\delta}_n)\right]^{-1}$$

Thus, the information is proportional to the sample size in all of the cases so far described. For example, in the single-sample case $I \propto n/\sigma^2$, in the two-sample comparison $I \propto n/\sigma^2_\Delta$ and in the single-proportion case $I \propto \frac{n}{p(1-p)}$.

However, the statistical information in time-to-event trials that are based on the log-rank test is $I \propto D$, i.e., it is not proportional to the number of subjects but the number of events!

Statistical information
Studies of time to event

In the case of time to event studies, there are a number of considerations with respect to study design. These are:

- *Accrual of patients*.
  Accrual of patients happens at a rate of $a(t)$ over time.
- *Follow-up of patients*.
  Follow-up of patients happens over time $t-u$ after they have been accrued at time u.
- *Hazard* $\lambda(t)$.
  Hazard time $\lambda(t) = \lim_{h\to 0} \frac{1}{h}\text{Pr}(t \le T \le T+h|T \ge t)$, is the instantaneous failure rate.
- *Survival distribution* $S(t)$
  Survival distribution function is $S(t) = P(T > t)$ is the probability of survival past time t.

Time-to-event studies
Event rate

In order to have an event, each individual participating in the study must

- Have been accrued at time $u < t$
- Survived during the period $t - u$
- Had an event at time t

The event rate at time $t \leq T$ is given in general by the expression

$$n(t) = \int_0^\tau \underbrace{a(u)}_{\text{accrued at } u} \underbrace{S(t-u)}_{\text{survive past } t-u} \underbrace{\lambda(t-u)}_{\text{fail at } t-u+h} du$$

where $\tau = \min(t, T)$, so that

$$n(t) = \begin{cases} \int_0^t a(u)S(t-u)\lambda(t-u)du & \text{if } t \leq T \\ \int_0^T a(u)S(t-u)\lambda(t-u)du & \text{if } t > T \end{cases}$$

Time-to-event studies
Simplifying assumptions

Accrual is usually assumed to be uniform over the period $[0, T]$, i.e.,

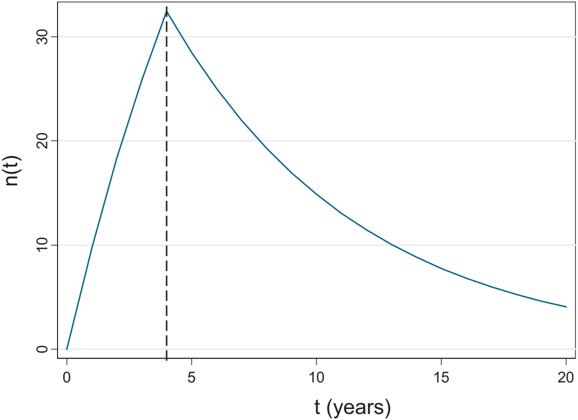
$$a(t) = \begin{cases} a_0 & \text{if } t \leq T, \\ 0 & \text{if } t > T \end{cases}$$

If the additional assumption is made that survival is exponential is made (i.e., $S(t) = \int_t^\infty \lambda e^{-\lambda u} du = e^{-\lambda t}$ and $\lambda(t) = \lambda$), the event rate at time t is

$$n(t) = \begin{cases} a_0 \int_0^t \lambda e^{-\lambda(t-u)} du = a_0(1 - e^{-\lambda t}) & \text{if } t \leq T, \\ a_0 \int_0^T \lambda e^{-\lambda(t-u)} du = a_0(e^{-\lambda(t-T)} - e^{-\lambda t}) & \text{if } t > T. \end{cases}$$

Time-to-event studies
Example (see Piantadosi, 2005, pp. 321-322)

Supposed that a clinical trial requires 180 events to achieve its planned power. If accrual proceeds at $a_0 = 80$ subjects annually, for $T = 4$ years, the event rate is constant at $\lambda = 0.13$ deaths per person-year of follow-up, the number of events is given graphically in the Figure below.



Time-to-event studies
Cumulative number of events

The number of total events is

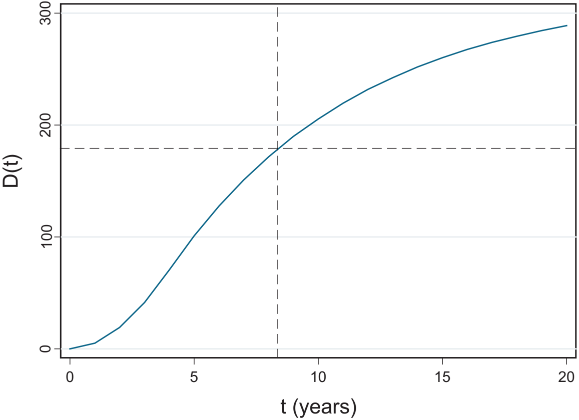
$$D(t) = \int_0^t n(u)du$$

Using the previous simplifying assumptions of a uniform accrual and exponential survival, the cummulative number of events is

$$\begin{array}{ll} \int_0^t a_0(1 - e^{-\lambda t})dt & \\ \qquad = \frac{a_0}{\lambda}(\lambda t + e^{-\lambda t} - 1) & \text{if } t \leq T \\ \int_0^T a_0(1 - e^{-\lambda t})dt + \int_T^t a_0(e^{-\lambda(t-T)} - e^{-\lambda t})dt & \\ \qquad = \frac{a_0}{\lambda}(\lambda T + e^{-\lambda t} - e^{-\lambda(t-T)}) & \text{if } t > T \end{array}$$

Clinical trial example
Cumulative number of events

In the previous example, the cumulative number of events is given graphically in the Figure below. Notice that it would take over five years after completion of accrual to reach the desired 180 number of events.



Designing survival studies
Number of events versus number of subjects

Designing studies with time to event as the endpoint, is challenging because we would like to, ultimately, determine the size of the sample. Given the complexities of the design, there is no single sample size that will fit the desired power considerations. So extensive experimentation is necessary. Here are some guidelines:

- If the cost of patient accrual is small relatively to the cost of a long study and accrual rates are fixed, we can accelerate the completion of the study by increasing the sample size (i.e., extending accrual versus extending follow-up). Alternatively, if we can, accrual rates can be increased by incorporating more sites.
- If the cost of patient accrual is great, we can accrue less patients (although never of course less than the desired number of events) and follow them longer (so that we can observe a larger proportion of them having the event)