Διακριτές τυχαίες μεταβλητές
Περίληψη

Αντώνης Οικονόμου
aeconom@math.uoa.gr

30 Μαρτίου 2010

Μια **τυχαία μεταβλητή** που αναφέρεται σε ένα πείραμα τύχης είναι διαισθητικά ένα αριθμητικό χαρακτηριστικό του πειράματος. Σε μαθηματικό - τυπικό επίπεδο είναι μια συνάρτηση με πεδίο ορισμού το δειγματικό χώρο και τιμές στο σύνολο των πραγματικών αριθμών. Αν X είναι μια τυχαία μεταβλητή, η συνάρτηση κατανομής της, $F(x)$, ορίζεται ως:

$$F(x) = P(X \leq x).$$

Όλοι οι υπολογισμοί πιθανοτήτων που αφορούν τη X μπορούν να γίνουν μέσω της συνάρτησης κατανομής.

Μια τυχαία μεταβλητή με πεπερασμένο ή το πολύ αριθμήσιμο πεδίο τιμών λέγεται **διακριτή** και αντιστοιχεί διαισθητικά σε κάποιο διακριτό αριθμητικό χαρακτηριστικό ενός πειράματος τύχης. Αν X είναι μια διακριτή τυχαία μεταβλητή, τότε η συνάρτηση

$$p(x) = P(X = x)$$

λέγεται **συνάρτηση πιθανότητας** της X. Η ποσότητα

$$E[X] = \sum_x xp(x)$$

λέγεται **μέση τιμή** της X. Οι όροι αναμενόμενη τιμή καθώς και μαθηματική ελπίδα της X χρησιμοποιούνται εναλλακτικά για την $E[X]$. Η $E[X]$ είναι ένα μέτρο θέσης της X, δηλαδή περιγράφει γύρω από ποιον αριθμό περιστρέφονται οι τιμές της X. Για να έχουμε μια πιο ικανοποιητική περιγραφή της συμπεριφοράς της X χρειαζόμαστε και κάποιο μέτρο συγκεντρωτικότητας ή μεταβλητότητας. Ως τέτοιο χρησιμοποιείται η ποσότητα

$$Var[X] = E[(X - E[X])^2]$$

που αναφέρεται ως **διασπορά** της X. Η $Var[X]$ περιγράφει πόσο διεσπαρμένες είναι οι τιμές της X γύρω από τη μέση τιμή της. Η ποσότητα $\sqrt{Var[X]}$ αναφέρεται ως τυπική απόκλιση της X.

Η μέση τιμή $E[g(X)]$ μιας συνάρτησης τυχαίας μεταβλητής μπορεί να βρεθεί άμεσα από τον τύπο

$$E[g(X)] = \sum_x g(x)p(x).$$

Χρήσιμοι τύποι για τον υπολογισμό της μέσης τιμής και της διασποράς μιας τυχαίας μεταβλητής είναι οι εξής:

$$\begin{aligned} E[aX+b] &= aE[X]+b, \\ Var[aX+b] &= a^2Var[X], \\ Var[X] &= E[X^2]-E[X]^2. \end{aligned}$$

Ένα σημαντικό πείραμα τύχης είναι η τυχαία επιλογή αριθμού από το σύνολο $\{1, 2, \ldots, N\}$. Αν X είναι το αποτέλεσμα του πειράματος, τότε λέμε ότι η X ακολουθεί την **διακριτή ομοιόμορφη κατανομή** στο $\{1, 2, \ldots, N\}$ (Uniform($\{1, 2, \ldots, N\}$)). Η συνάρτηση πιθανότητας της X δίνεται τότε ως

$$p(i) = \frac{1}{N}, \quad i = 1, 2, \ldots, N.$$

Η μέση τιμή και η διασπορά της είναι

$$E[X] = \frac{N+1}{2}, \quad Var[X] = \frac{N^2-1}{12}.$$

Ένα άλλο σημαντικό πείραμα τύχης είναι οι επαναλαμβανόμενες πραγματοποιήσεις ενός πειράματος με δυο πιθανές εκβάσεις, επιτυχία και αποτυχία, με πιθανότητα επιτυχίας p και πιθανότητα αποτυχίας $1-p$. Αυτές αναφέρονται και ως ανεξάρτητες δοκιμές Bernoulli με πιθανότητα επιτυχίας p. Αν X είναι ο αριθμός των επιτυχιών σε n ανεξάρτητες δοκιμές Bernoulli, τότε λέμε ότι η X ακολουθεί τη **διωνυμική κατανομή** με παραμέτρους n και p (Bin(n, p)). Η συνάρτηση πιθανότητας της X δίνεται τότε ως

$$p(i) = \binom{n}{i} p^i (1-p)^{n-i}, \quad i = 0, 1, 2, \ldots, n.$$

Η μέση τιμή και η διασπορά της είναι

$$E[X] = np, \quad Var[X] = np(1-p).$$

Αν X είναι ο αριθμός των δοκιμών μέχρι την πρώτη επιτυχία σε μια σειρά ανεξάρτητων δοκιμών Bernoulli, τότε λέμε ότι η X ακολουθεί τη **γεωμετρική κατανομή** με παράμετρο p στο σύνολο $\{1, 2, \ldots\}$ (Geom(p) στο $\{1, 2, \ldots\}$). Η συνάρτηση πιθανότητας της X δίνεται τότε ως

$$p(i) = p(1-p)^{i-1}, \quad i = 1, 2, \ldots.$$

Η μέση τιμή και η διασπορά της είναι

$$E[X] = \frac{1}{p}, \quad Var[X] = \frac{1-p}{p^2}.$$

Αν X είναι ο αριθμός των δοκιμών μέχρι την r-οστή επιτυχία σε μια σειρά ανεξάρτητων δοκιμών Bernoulli, τότε λέμε ότι η X ακολουθεί την **αρνητική διωνυμική κατανομή** με παραμέτρους r και p στο σύνολο $\{r, r+1, \ldots\}$ (NegBin(r, p) στο $\{r, r+1, \ldots\}$). Η συνάρτηση πιθανότητας της X δίνεται τότε ως

$$p(i) = \binom{i-1}{r-1} p^r (1-p)^{i-r}, \quad i = r, r+1, \ldots.$$

Η μέση τιμή και η διασπορά της είναι

$$E[X] = \frac{r}{p}, \quad Var[X] = \frac{r(1-p)}{p^2}.$$

Στην περίπτωση που έχουμε έναν μεγάλο αριθμό n ανεξάρτητων δοκιμών Bernoulli και η πιθανότητα επιτυχίας p σε κάθε δοκιμή είναι μικρή, έτσι ώστε $np = \lambda$, τότε ο αριθμός των επιτυχιών προσεγγίζεται από τη λεγόμενη **κατανομή Poisson** με παράμετρο λ (Poisson(λ)). Η συνάρτηση πιθανότητας της X δίνεται τότε ως

$$p(i) = e^{-\lambda} \frac{\lambda^i}{i!}, \quad i = 0, 1, \ldots.$$

Η μέση τιμή και η διασπορά της είναι

$$E[X] = \lambda, \quad Var[X] = \lambda.$$

Η διωνυμική κατανομή έχει και την εξής εναλλακτική ερμηνεία: Υποθέτουμε ότι σε έναν (άπειρο) πληθυσμό, ένα ποσοστό p των ατόμων του πληθυσμού έχει κάποιο χαρακτηριστικό (επιτυχία) ενώ το ποσοστό $1-p$ των ατόμων του πληθυσμού δεν έχει το χαρακτηριστικό (αποτυχία). Αν επιλέξουμε n άτομα του πληθυσμού και θέσουμε X την τυχαία μεταβλητή που παριστάνει το πλήθος των ατόμων που έχουν το χαρακτηριστικό, τότε η X ακολουθεί τη διωνυμική κατανομή με παραμέτρους n και p. Στην περίπτωση πεπερασμένου πληθυσμού N ατόμων από τα οποία τα m έχουν το χαρακτηριστικό ενώ τα $N-m$ δεν έχουν το χαρακτηριστικό, αν επιλέξουμε n άτομα του πληθυσμού και θέσουμε X το πλήθος των ατόμων που έχουν το χαρακτηριστικό, τότε η κατανομή της X εξαρτάται από το κατά πόσον η δειγματοληψία γίνεται με ή χωρίς επανάθεση.

Αν η δειγματοληψία γίνεται με επανάθεση τότε η κατάσταση δεν διαφέρει από τον άπειρο πληθυσμό και έχουμε ότι η X ακολουθεί τη διωνυμική κατανομή με παραμέτρους n και $\frac{m}{N}$. Αν, όμως, η δειγματοληψία γίνεται χωρίς επανάθεση τότε η X έχει συνάρτηση πιθανότητας

$$p(i) = \frac{\binom{m}{i}\binom{N-m}{n-i}}{\binom{N}{n}}, \quad i = 0, 1, \ldots, n.$$

Αν θέσουμε $p = \frac{m}{N}$ τότε η μέση τιμή και η διασπορά της είναι

$$E[X] = np, \quad Var[X] = \frac{N-n}{N-1}np(1-p).$$

Η X λέμε τότε ότι ακολουθεί την **υπεγεωμετρική κατανομή** με παραμέτρους n, N, m ($\mathrm{Hypergeom}(n, N, m)$). Για μεγάλα N και σχετικά μικρά n η υπεργεωμετρική κατανομή $\mathrm{Hypergeom}(n, N, m)$ προσεγγίζεται ικανοποιητικά από την διωνυμική κατανομή $\mathrm{Bin}(n, p)$ με $p = \frac{m}{N}$. Αυτό είναι φανερό διότι εάν έχουμε έναν μεγάλο πληθυσμό (δηλαδή μεγάλο N) και πρόκειται να πάρουμε μικρό δείγμα (δηλαδή μικρό n) η κατανομή του πλήθους των ατόμων που έχουν το χαρακτηριστικό αναμένουμε να είναι περίπου η ίδια, είτε έχουμε δειγματοληψία με επανάθεση, είτε δειγματοληψία χωρίς επανάθεση.