Εισαγωγή στις (διακριτές) τυχαίες μεταβλητές

Πιθανότητες και Στατιστική

Τμήμα Πληροφορικής, Πανεπιστήμιο Πειραιώς

2023-2024

Παράδειγμα

Θεωρούμε το εξής πείραμα: Ρίχνουμε ένα νόμισμα 4 φορές. Συμβολίζουμε με

- X τον αριθμό των εμφανίσεων της όψης ΓΡΑΜΜΑΤΑ.
- Y τον αριθμό της ρίψης που εμφανίσθηκε πρώτη φορά η όψη ΓΡΑΜΜΑΤΑ.
- Z τη διαφορά του αριθμού των εμφανίσεων της όψης ΓΡΑΜΜΑΤΑ και του αριθμού των εμφανίσεων της όψης ΚΟΡΩΝΑ, δηλαδή $Z = X - (4 - X) = 2X - 4$.

Οι τιμές των X, Y, Z δεν είναι γνωστές εκ των προτέρων, αλλά εξαρτώνται από την έκβαση του πειράματος.

Οι X, Y, Z αποτελούν παραδείγματα τυχαίων μεταβλητών. Διαισθητικά, μια τυχαία μεταβλητή είναι μια ποσότητα που εξαρτάται από την έκβαση ενός τυχαίου πειράματος.

Τυχαίες Μεταβλητές - Εισαγωγή

Ορισμός

Έστω Ω ένας δειγματικός χώρος. Αν σε κάθε στοιχειώδες ενδεχόμενο $\{\omega\}$ του Ω αντιστοιχίσουμε ένα πραγματικό αριθμό, τότε ορίζεται μια συνάρτηση $X : \Omega \to \mathbb{R}$, η οποία ονομάζεται **τυχαία μεταβλητή** (ΤΜ). (Συνήθως, οι ΤΜ συμβολίζονται με κεφαλαία γράμματα.)
Το **σύνολο τιμών** (range) της X συμβολίζεται με S_X ή απλά με S, δηλαδή $S_X := X(\Omega)$.
Αν το S_X είναι το πολύ αριθμήσιμο, τότε η X ονομάζεται **διακριτή** ΤΜ.

Τυχαίες Μεταβλητές - Εισαγωγή

Το ενδεχόμενο μια ΤΜ X να λάβει την τιμή $x \in S_X$ συμβολίζεται με $\{X = x\}$, δηλαδή

$$\{X = x\} := X^{-1}(x) := \{\omega \in \Omega : X(\omega) = x\},$$

και η πιθανότητα του ενδεχομένου αυτού με

$$P(X = x) := P(\{\omega \in \Omega : X(\omega) = x\}).$$

(Η πιθανότητα $P(X = x)$ ορίζεται βάσει του μέτρου πιθανότητας που έχει ήδη ορισθεί στο Ω.)
Πιο γενικά, για κάθε $T \subseteq S_X$, συμβολίζουμε με $\{X \in T\}$ το ενδεχόμενο $\bigcup_{x \in T}\{X = x\}$, δηλαδή είναι

$$\{X \in T\} := X^{-1}(T) := \{\omega \in \Omega : X(\omega) \in T\},$$

και την αντίστοιχη πιθανότητα με

$$P(X \in T) := P(\{\omega \in \Omega : X(\omega) \in T\}).$$

Οι ΤΜ X, Y, Z του παραδείγματος λαμβάνουν τις παρακάτω τιμές, για κάθε στοιχειώδες ενδεχόμενο του δειγματικού χώρου Ω (πρώτη στήλη του πίνακα):

$\omega \in \Omega$	X	Y	Z
ΚΚΚΚ	0	0	-4
ΚΚΚΓ	1	4	-2
ΚΚΓΚ	1	3	-2
ΚΚΓΓ	2	3	0
ΚΓΚΚ	1	2	-2
ΚΓΚΓ	2	2	0
ΚΓΓΚ	2	2	0
ΚΓΓΓ	3	2	2

$\omega \in \Omega$	X	Y	Z
ΓΚΚΚ	1	1	-2
ΓΚΚΓ	2	1	0
ΓΚΓΚ	2	1	0
ΓΚΓΓ	3	1	2
ΓΓΚΚ	2	1	0
ΓΓΚΓ	3	1	2
ΓΓΓΚ	3	1	2
ΓΓΓΓ	4	1	4

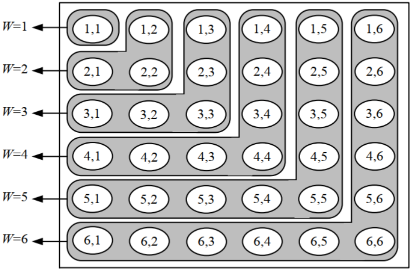
Το σύνολο τιμών (δυνατές τιμές) της X είναι $S_X = \{0, 1, 2, 3, 4\}$.
Το σύνολο τιμών της Y είναι $S_Y = \{0, 1, 2, 3, 4\}$.
Το σύνολο τιμών της Z είναι $S_Z = \{-4, -2, 0, 2, 4\}$.

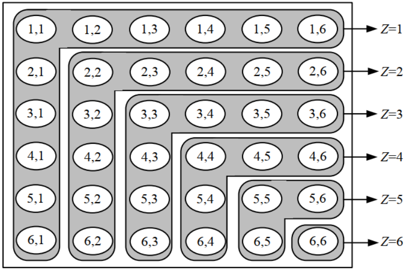
Η X λαμβάνει την τιμή 0 αν δεν εμφανίσθηκε καθόλου η όψη ΓΡΑΜΜΑΤΑ, δηλαδή μόνο για το ενδεχόμενο ΚΚΚΚ, το οποίο εμφανίζεται με πιθανότητα $\frac{1}{2^4}$.
Η X λαμβάνει την τιμή 1 αν εμφανίσθηκε ακριβώς μια όψη ΓΡΑΜΜΑΤΑ, δηλαδή για τα ενδεχόμενα ΓΚΚΚ, ΚΓΚΚ, ΚΚΓΚ, ΚΚΚΓ, καθένα εκ των οποίων εμφανίζεται με πιθανότητα $\frac{1}{16}$, άρα η X λαμβάνει την τιμή 1 με πιθανότητα $\frac{4}{16} = \frac{1}{4}$.

Παράδειγμα

Ρίχνουμε δύο ζάρια και συμβολίζουμε με X, Y τα αντίστοιχα αποτελέσματα των 2 ρίψεων.
Ο δειγματικός χώρος είναι ο $\Omega = \{(x, y) : x, y \in [6]\}$.
Προφανώς οι X, Y είναι ΤΜ, με $S_X = S_Y = [6]$.
Επιπλέον ορίζουμε τις ΤΜ $W = \max\{X, Y\}$ και $Z = \min\{X, Y\}$.
Προφανώς, είναι $S_W = S_Z = [6]$.
Τα ενδεχόμενα $\{W = k\}$ και $\{Z = k\}$ αναπαριστάνονται στο επόμενο σχήμα (ως υποσύνολα του Ω) για κάθε $k \in [6]$.

Τυχαίες Μεταβλητές - Εισαγωγή





Παρατηρήστε ότι η οικογένεια $(\{W = k\})_{k\in[6]}$ διαμερίζει το Ω, όπως και η οικογένεια $(\{Z = k\})_{k\in[6]}$.
Με συμβολισμό συναρτήσεων, οι διαμερίσεις αυτές είναι οι $(W^{-1}(k))_{k\in[6]}$ και $(Z^{-1}(k))_{k\in[6]}$ αντίστοιχα.

Συνάρτηση μάζας πιθανότητας

Ορισμός

Έστω μια διακριτή ΤΜ X με σύνολο τιμών S. Η συνάρτηση $f_X(x) = P(X = x)$ με πεδίο ορισμού το S, ονομάζεται **συνάρτηση μάζας πιθανότητας** (probability mass function - PMF) ή **μάζα** της ΤΜ X. (Συνήθως γράφουμε $f(x)$ αντί για $f_X(x)$ όταν δεν υπάρχει κίνδυνος σύγχυσης.)

Παράδειγμα

Να βρεθούν οι PMF των ΤΜ X, Y, Z του πρώτου παραδείγματος, όπου

- X ο αριθμός εμφανίσεων της όψης ΓΡΑΜΜΑΤΑ,
- Y ο αριθμός της ρίψης που εμφανίσθηκε πρώτη φορά ΓΡΑΜΜΑΤΑ,
- Z η διαφορά των εμφανίσεων ΓΡΑΜΜΑΤΑ μείον τις εμφανίσεις ΚΟΡΩΝΑ, δηλαδή $Z = X - (4 - X) = 2X - 4$,

μετά από 4 ρίψεις ενός νομίσματος.

Λύση

Για την X, είναι $S_X = \{0,1,2,3,4\}$ και

$$f_X(0) = P(X=0) = \frac{1}{16},\ f_X(1) = P(X=1) = \frac{4}{16},\ f_X(2) = P(X=2) = \frac{6}{16},$$

$$f_X(3) = P(X=3) = \frac{4}{16}, \quad f_X(4) = P(X=4) = \frac{1}{16}.$$

Για την Y, είναι $S_Y = \{0,1,2,3,4\}$, $f_Y(0) = P(Y=0) = 1/16$ και

$$x \in [4] \Rightarrow f_Y(x) = P(Y=x) = \frac{1}{2^{x-1}}\frac{1}{2} = 2^{-x}.$$

Για την Z, είναι $S_Z = \{-4,-2,0,2,4\}$ και για κάθε $x \in S_Z$

$$f_Z(x) = P(Z=x) = P(2X-4=x) = P(X=2+x/2) = f_X(2+x/2).$$

Συνάρτηση μάζας πιθανότητας

Πρόταση

Η PMF f_X μιας διακριτής ΤΜ X με σύνολο τιμών S_X έχει τις εξής ιδιότητες:

- $0 \leq f_X(x) \leq 1$, για κάθε $x \in S_X$.
- $\sum_{x \in S_X} f_X(x) = 1$.
- $P(X \in T) = \sum_{x \in T} f_X(x)$, για κάθε $T \subseteq S_X$.

Απόδειξη.

Η πρώτη ανισότητα είναι προφανής, αφού εξ ορισμού είναι $f_X(x) = P(X = x)$ και $P(X = x) \in [0, 1]$.
Για τη δεύτερη ιδιότητα, παρατηρούμε ότι η οικογένεια $(X^{-1}(x))_{x \in S_X}$ αποτελεί διαμέριση του Ω, επομένως

$$\sum_{x \in S_X} f_X(x) = \sum_{x \in S_X} P(X = x) = \sum_{x \in S_X} P(X^{-1}(x)) = P(\Omega) = 1.$$

Η απόδειξη της τρίτης ιδιότητας είναι ανάλογη (άσκηση). $\square$

Συνάρτηση μάζας πιθανότητας

Το επόμενο αποτέλεσμα αποτελεί τον τύπο ολικής πιθανότητας για διακριτές τυχαίες μεταβλητές.

Λήμμα

Αν X, Y διακριτές ΤΜ, τότε

$$P(X = x) = \sum_{y \in S_Y} P(X = x, Y = y), \qquad \text{για κάθε } x \in S_X.$$

Απόδειξη.

Το ενδεχόμενο $A = \{X = x\}$ διαμερίζεται στα $A_y = A \cap B_y$, $y \in S_Y$, όπου $B_y = \{Y = y\}$. Επομένως,

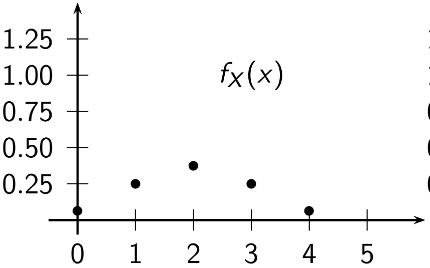
$$P(X = x) = P(A) = \sum_{y \in S_Y} P(A \cap B_y) = \sum_{y \in S_Y} P(X = x, Y = y).$$

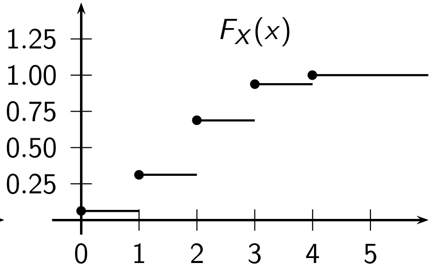
$\square$

Συνάρτηση κατανομής

Ορισμός

Η (αθροιστική) συνάρτηση κατανομής (cumulative distribution function - CDF) της ΤΜ X είναι η συνάρτηση $F_X : \mathbb{R} \to [0,1]$, με $F_X(x) = P(X \leq x)$ για κάθε $x \in \mathbb{R}$,





Σχήμα: Η συναρτήσεις f_X και F_X για την ΤΜ X του προηγούμενου παραδείγματος.

Για την (αθροιστική) συνάρτηση κατανομής πιθανότητας $F(x)$ ισχύουν οι παρακάτω ιδιότητες

1. Η $F(x)$ είναι αύξουσα.
2. Η $F(x)$ είναι συνεχής από τα δεξιά, δηλαδή $\lim_{x \to x_0^+} F(x) = F(x_0)$ για κάθε $x_0 \in \mathbb{R}$.
3. $\lim_{x \to \infty} F(x) = 1$ και $\lim_{x \to -\infty} F(x) = 0$.
4. $P(X < x_0) = \lim_{x \to x_0^-} F(x)$.
5. $P(X > x) = 1 - P(X \leq x) = 1 - F(x)$.
6. $P(x_1 < X \leq x_2) = P(X \leq x_2) - P(X \leq x_1) = F(x_2) - F(x_1)$.

Ορισμός

Η **μέση** (ή προσδοκώμενη, ή αναμενόμενη) **τιμή** (expected value) μιας διακριτής ΤΜ X με PMF f_X ορίζεται ως ο αριθμός $E(X)$ με

$$E(X) := \sum_{x \in S_X} xP(X = x) = \sum_{x \in S_X} x f_X(x).$$

Γενικότερα, για οποιαδήποτε συνάρτηση $g : S_X \to \mathbb{R}$, η μέση τιμή της νέας ΤΜ $g(X)$ δίνεται από τη σχέση

$$E(g(X)) = \sum_{x \in S_X} g(x)P(X = x) = \sum_{x \in S_X} g(x) f_X(x).$$

Παρατήρηση: Η μέση τιμή μιας ΤΜ μπορεί και να απειρίζεται. Στο εξής, θα θεωρούμε πάντα ότι η μέση τιμή είναι πεπερασμένη, εκτός αν αναφέρεται ρητά το αντίθετο.

Ο δεύτερος τύπος είναι άμεση συνέπεια του πρώτου. Πράγματι, η ΤΜ $Y = g(X)$ ορίζεται στον (το πολύ αριθμήσιμο) δειγματικό χώρο S_X, ο οποίος διαμερίζεται από την οικογένεια $(g^{-1}(y))_{y \in S_y}$ και έχει μέτρο πιθανότητας την f_X, επομένως

$$
\begin{aligned}
E(Y) &= \sum_{y \in S_Y} yP(Y = y) = \sum_{y \in S_Y} y \sum_{x \in g^{-1}(y)} f_X(x) \\
&= \sum_{y \in S_Y} \sum_{x \in g^{-1}(y)} g(x) f_X(x) = \sum_{x \in S_X} g(x) f_X(x).
\end{aligned}
$$

Αναμενόμενη τιμή και διακύμανση

Ορισμός

Η **διακύμανση** (ή διασπορά) (variance) μιας διακριτής ΤΜ X, με $E(X) = \mu \in \mathbb{R}$, ορίζεται ως ο αριθμός $\sigma^2 = V(X)$ με

$$V(X) = E[(X - \mu)^2].$$

Ο μη αρνητικός αριθμός $\sigma = \sqrt{V(X)}$ ονομάζεται **τυπική απόκλιση** (standard deviation) της X.

Παρατηρήσεις: Παρατηρήστε ότι η διακύμανση της X είναι ουσιαστικά η μέση τιμή της ΤΜ

$$Y = (X - E(X))^2.$$

Επομένως, η διακύμανση είναι ένα μέτρο που μας δείχνει πόσο συγκεντρωμένες βρίσκονται οι τιμές μιας ΤΜ γύρω από την μέση τιμή της. Αφού η Y παίρνει μη αρνητικές τιμές, η μέση τιμή της θα είναι προφανώς μη αρνητική, οπότε η τυπική απόκλιση σ ορίζεται πάντα.

Αναμενόμενη τιμή και διακύμανση

Πρόταση

Για οποιεσδήποτε διακριτές ΤΜ X, Y και σταθερές $a, b \in \mathbb{R}$, ισχύουν οι παρακάτω ιδιότητες:

i) $E(aX + bY) = aE(X) + bE(Y)$

ii) $V(aX + b) = a^2 V(X)$.

iii) $V(X) = E(X^2) - (E(X))^2$.

Απόδειξη.

i) Έστω η ΤΜ $Z = aX + bY$. Τότε είναι
$S_Z = \{ax + by : x \in S_X, y \in S_Y\}$ και $P(Z = z) = P(X = x, Y = y)$, για κάθε $z = ax + by \in S_Z$. Επομένως, βάσει του Λήμματος 12, προκύπτει ότι

$$
\begin{aligned}
E[Z] &= \sum_{z \in S_Z} zP(Z = z) = \sum_{x \in S_X} \sum_{y \in S_Y} (ax + by)P(X = x, Y = y) \\
&= a \sum_{x \in S_X} x \sum_{y \in S_Y} P(X = x, Y = y) + b \sum_{y \in S_Y} y \sum_{x \in S_X} P(X = x, Y = y) \\
&= a \sum_{x \in S_X} xP(X = x) + b \sum_{y \in S_Y} yP(Y = y) = aE(X) + bE(Y).
\end{aligned}
$$

Απόδειξη (συνέχεια).

ii) Βάσει του προηγούμενου αποτελέσματος, ισχύει ότι $E(aX+b) = aE(X) + b = a\mu + b$, οπότε

$$\begin{aligned} V(aX+b) &= E((aX+b-(a\mu+b))^2) \\ &= E(a^2(X-\mu)^2) = a^2E((X-\mu)^2) \\ &= a^2V(X). \end{aligned}$$

iii) Βάσει των προηγουμένων,

$$\begin{aligned} V(X) &= E((X-\mu)^2) = E(X^2 - 2\mu X + \mu^2) \\ &= E(X^2) - 2\mu E(X) + \mu^2 = E(X^2) - 2\mu^2 + \mu^2 \\ &= E(X^2) - \mu^2. \end{aligned}$$

□

Αναμενόμενη τιμή και διακύμανση

Παράδειγμα - Δείκτρια συνάρτηση

Έστω ενδεχόμενο $A \subseteq \Omega$ και έστω η διακριτή ΤΜ I_A, με

$$I_A = \begin{cases} 1, & \text{αν πραγματοποιείται το } A, \\ 0, & \text{αλλιώς.} \end{cases}$$

Η ΤΜ ονομάζεται **δείκτρια** (indicator) του ενδεχομένου A και αποτελεί μια **δυαδική** ΤΜ, αφού παίρνει τιμές 0 και 1.
Παρατηρήστε ότι αφού η ΤΜ I_A είναι μια συνάρτηση $I_A : \Omega \to \mathbb{R}$, μπορούμε ισοδύναμα να την ορίσουμε ως

$$I_A(\omega) = \begin{cases} 1, & \omega \in A, \\ 0, & \omega \notin A, \end{cases}$$

δηλαδή πρόκειται για την χαρακτηριστική συνάρτηση του συνόλου A.
Αν θέσουμε $p = P(A)$, τότε, βάσει του ορισμού, η μάζα της I_A είναι μια συνάρτηση $f : \{0, 1\} \to [0, 1]$ με $f(0) = 1 - p$ και $f(1) = p$.

Παράδειγμα - Δείκτρια συνάρτηση (συνέχεια)

Η μέση τιμή της ισούται με

$$E(I_A) = \sum_{x=0}^{1} x f(x) = 0 \cdot (1-p) + 1 \cdot p = p.$$

Διαισθητικά, αν επαναλάβουμε το πείραμα πολλές φορές, το ποσοστό πραγματοποιήσεων του A θα είναι περίπου $p \cdot 100\%$.
Η διακύμανσή της ισούται με

$$\begin{aligned} V(I_A) = E[(I_A - p)^2] &= \sum_{x=0}^{1} (x-p)^2 f(x) = p^2(1-p) + (1-p)^2 p \\ &= p(1-p) = p - p^2. \end{aligned}$$

Η παράσταση αυτή ελαχιστοποιείται όταν $p \in \{0, 1\}$ και μεγιστοποιείται όταν $p = 1/2$. Διαισθητικά, το αποτέλεσμα γίνεται λιγότερο προβλέψιμο, όταν το p πλησιάζει το $1/2$.

Αναμενόμενη τιμή και διακύμανση

Πόρισμα

Αν $X_1, X_2, \ldots, X_n$ διακριτές τυχαίες μεταβλητές και $g_1, g_2, \ldots, g_n$ οποιεσδήποτε συναρτήσεις, όπου $g_i : S_{X_i} \to \mathbb{R}$, τότε:

$$E(g_1(X_1)+g_2(X_2)+\cdots+g_n(X_n)) = E(g_1(X_1))+E(g_2(X_2))+\cdots+E(g_n(X_n)).$$

Ο τύπος του προηγούμενου πορίσματος είναι ιδιαίτερα χρήσιμος για την απλούστευση του υπολογισμού μέσης τιμής σε πολλές εφαρμογές. Αντίστοιχος τύπος ισχύει και για την διακύμανση, μόνο στην περίπτωση που οι τυχαίες μεταβλητές είναι ανεξάρτητες.

Ορισμός

Οι τυχαίες μεταβλητές $X_1, X_2, \ldots, X_n$ καλούνται ανεξάρτητες όταν

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = P(X_1 = x_1)P(X_2 = x_2)\cdots P(X_n = x_n),$$

για όλες τις δυνατές τιμές των $x_1, x_2, \ldots, x_n$.

Αναμενόμενη τιμή και διακύμανση

Πρόταση

Αν οι τυχαίες μεταβλητές $X_1, X_2, \ldots, X_n$ είναι ανεξάρτητες και $g_1, g_2, \ldots, g_n$ οποιεσδήποτε συναρτήσεις, όπου $g_i : S_{X_i} \to \mathbb{R}$, τότε:

i) Οι $g_1(X_1), g_2(X_2), \ldots, g_n(X_n)$ είναι ανεξάρτητες

ii) $E(g_1(X_1)g_2(X_2)\cdots g_n(X_n)) = E(g_1(X_1))E(g_2(X_2))\cdots E(g_n(X_n))$

iii) $V\left(\sum_{i=1}^{n} g_i(X_i)\right) = \sum_{i=1}^{n} V\left(g_i(X_i)\right)$

Απόδειξη.

i) Αρχικά, θα δειχθεί ότι $P(X_1 \in A_1, X_2 \in A_2) = P(X_1 \in A_1)P(X_2 \in A_2)$, για οποιαδήποτε $A_1 \subseteq S_{X_1}$, $A_2 \subseteq S_{X_2}$. Πράγματι, είναι

$$\begin{aligned} P(X_1 \in A_1, X_2 \in A_2) &= \sum_{(x,y)\in A_1\times A_2} P(X_1 = x, X_2 = y) = \sum_{(x,y)\in A_1\times A_2} P(X_1 = x)P(X_2 = y) \\ &= \sum_{x\in A_1} P(X_1 = x) \sum_{y\in A_2} P(X_2 = y) = P(X \in A_1)P(X \in A_2) \end{aligned}$$

Επομένως, για κάθε $(x, y) \in \mathbb{R}^2$ είναι

$$\begin{aligned} P(g_1(X_1) = x, g_2(X_2) = y) &= P(X_1 \in g_1^{-1}(x), X_2 \in g_2^{-1}(y)) \\ &= P(X_1 \in g_1^{-1}(x))P(X_2 \in g_2^{-1}(y)) \\ &= P(g_1(X_1) = x)P(g_2(X_2) = y) \end{aligned}$$

δηλαδή οι $g_1(X_1)$, $g_2(X_2)$ είναι ανεξάρτητες. Το αποτέλεσμα αυτό γενικεύεται επαγωγικά για n όρους.

Απόδειξη (συνέχεια).

ii) Ομοίως με την Πρόταση 18 (1). ('σκηση.)
iii)Θέτουμε $E(X_1) = \mu_1$ και $E(X_2) = \mu_2$, οπότε $E(X_1X_2) = \mu_1\mu_2$, $E(X_1 + X_2) = \mu_1 + \mu_2$ και

$$\begin{aligned}
V(X_1+X_2) &= E((X_1 + X_2 - \mu_1 - \mu_2)^2) \\
&= E((X_1-\mu_1)^2 + (X_2-\mu_2)^2 - 2(X_1-\mu_1)(X_2-\mu_2)) \\
&= E((X_1-\mu_1)^2) + E((X_2-\mu_2)^2) - 2E((X_1-\mu_1)(X_2-\mu_2)) \\
&= V(X_1) + V(X_2) - 2E(X_1 - \mu_1)E(X_2 - \mu_2) \\
&= V(X_1) + V(X_2) - 2(E(X_1) - \mu_1)(E(X_2) - \mu_2) \\
&= V(X_1) + V(X_2).
\end{aligned}$$

Κατόπιν, το αποτέλεσμα αυτό γενικεύεται επαγωγικά για *n* όρους.

Αναμενόμενη τιμή και διακύμανση

Παράδειγμα

Ρίχνουμε ένα κέρμα n φορές και θέτουμε X το πλήθος των φορών που εμφανίσθηκε κορώνα. Να υπολογισθούν οι $E(X)$ και $V(X)$.

Λύση

Θεωρούμε την δυαδική ΤΜ X_i, $i \in [n]$, με $X_i = 1$ αν εμφανίσθηκε κορώνα κατά την i-οστή ρίψη, και $X_i = 0$ αλλιώς. Με άλλα λόγια, η X_i είναι η δείκτρια του ενδεχομένου A_i να έρθει κορώνα κατά την i-οστή ρίψη. Θεωρούμε ότι $P(A_i) = p$, για κάθε $i \in [n]$, οπότε, βάσει του προηγούμενου παραδείγματος, έχουμε ότι $E(X_i) = p$ και $V(X_i) = p(1-p)$. Προφανώς, είναι $X = X_1 + X_2 + \cdots + X_n$, οπότε

$$E(X) = E(X_1 + X_2 + \cdots + X_n) = E(X_1) + E(X_2) + \cdots + E(X_n) = np.$$

Επιπλέον, επειδή οι X_i είναι ανεξάρτητες (αφού οι ρίψεις είναι ανεξάρτητες), έχουμε ότι
$V(X) = V(X_1+X_2+\cdots+X_n) = V(X_1)+V(X_2)+\cdots+V(X_n) = np(1-p).$

Κωδικοποιώντας στο προηγούμενο παράδειγμα την κορώνα με 1 και τα γράμματα με 0, το προηγούμενο πείραμα ισοδυναμεί με την τυχαία επιλογή μιας δυαδικής λέξης με n γράμματα, όταν κάθε γράμμα είναι 1 με πιθανότητα p. Η ΤΜ X αντιστοιχεί στο πλήθος των 1 στη λέξη. Έστω Y το πλήθος των εμφανίσεων του μοτίβου 11 στη δυαδική λέξη (εμφάνιση κορώνας 2 συνεχόμενες φορές). Ποιά είναι η μέση τιμή και η διακύμανση της Y;
Για την απάντηση, θεωρούμε τη δείκτρια ΤΜ Y_i, $i \in [n-1]$, με $Y_i = 1$ ανν εμφανίζεται το μοτίβο 11 στις θέσεις i και $i+1$. Προφανώς, είναι $E(Y_i) = p^2$ και $V(Y_i) = p^2(1-p^2)$ (γιατί;). Επομένως,

$$E(Y) = E(Y_1+Y_2+\cdots+Y_{n-1}) = E(Y_1)+E(Y_2)+\cdots+E(Y_{n-1}) = (n-1)p^2.$$

Όμως, οι Y_i δεν είναι ανεξάρτητες, οπότε δεν μπορεί να χρησιμοποιηθεί ο αντίστοιχος τύπος για τη διακύμανση $V(Y)$.

Αναμενόμενη τιμή και διακύμανση

Πρόταση

Αν η X είναι ΤΜ με τιμές στο $\mathbb{N}$, τότε $E(X) = \sum_{n=1}^{\infty} P(X \geq n)$.

Απόδειξη.

Εφαρμόζοντας τον ορισμό για τη μέση τιμή της X και δεδομένου ότι $n = \sum_{i=1}^{n} 1$, προκύπτει ότι

$$
\begin{aligned}
E(X) &= \sum_{n=1}^{\infty} nP(X = n) = \sum_{n=1}^{\infty} \sum_{i=1}^{n} P(X = n) \\
&= \sum_{i=1}^{\infty} \sum_{n=i}^{\infty} P(X = n) = \sum_{i=1}^{\infty} P(X \geq i).
\end{aligned}
$$

□

Άσκηση

Σε ένα πρόγραμμα επιλογής τυχαίων φυσικών αριθμών από το 1 έως το n οι αριθμοί έχουν πιθανότητα εμφάνισης ανάλογη του μεγέθους τους, δηλαδή η πιθανότητα να εμφανισθεί ο αριθμός k ισούται με $C \cdot k$, για κάθε $k \in [n]$, όπου C είναι μια σταθερά.

i) Να βρεθεί η πιθανότητα εμφάνισης κάθε αριθμού k, όπου $k = 1, \ldots, n$, δηλαδή να προσδιορισθεί η σταθερά C. (Υπόδειξη: Το άθροισμα όλων των πιθανοτήτων ισούται με 1.)

ii) Προτείνετε έναν τρόπο υλοποίησης ενός προγράμματος παραγωγής τυχαίων φυσικών αριθμών στο σύνολο $[n]$ με τις παραπάνω πιθανότητες.

iii) Να υπολογισθεί η πιθανότητα εμφάνισης του αριθμού 1 δεδομένου ότι επιλέχθηκε αριθμός από το 1 έως το ν, όπου $\nu \leq n$.

iv) Να βρεθεί η αναμενόμενη τιμή των αριθμών που εμφανίζονται.

v) Να βρεθεί η διακύμανση των αριθμών που εμφανίζονται.

Λύση

Έστω X ο αριθμός που θα εμφανισθεί.
i) Η διακριτή ΤΜ X λαμβάνει τιμές στο σύνολο $S = \{1, 2, \ldots, n\}$ και η PMF της X ικανοποιεί τις σχέσεις

$$f(k) = P(X = k) = Ck, \text{ για κάθε } k \in [n].$$

Επιπλέον, πρέπει

$$\sum_{k=1}^{n} f(k) = 1$$

οπότε

$$C(1 + 2 + \cdots + n) = 1 \Leftrightarrow C = \frac{2}{n(n+1)}$$

Άρα,

$$f(k) = P(X = k) = \frac{2k}{n(n+1)}.$$

Λύση (συνέχεια)

ii) Ένας τρόπος υλοποίησης είναι να παράγουμε ομοιόμορφα φυσικούς αριθμούς στο σύνολο

$$A = \left\{1, \dots, \frac{n(n+1)}{2}\right\}.$$

Αν ο αριθμός x που παράγεται είναι στο σύνολο

$$A_k = \left\{1 + \frac{(k-1)k}{2}, \dots, \frac{k(k+1)}{2}\right\},$$

για κάποιο $k \in [n]$ τότε εξάγουμε τον αριθμό k.
Τα σύνολα A_k, $k \in [n]$ αποτελούν μια διαμέριση του A και το καθένα περιέχει k φυσικούς, οπότε η πιθανότητα να ανήκει ο x στο A_k είναι $\frac{2k}{n(n+1)}$, επομένως ο αριθμός k παράγεται με τη ζητούμενη πιθανότητα.

Λύση (συνέχεια)

iii) Η ζητούμενη πιθανότητα ισούται με

$$\begin{aligned}P(X = 1|X \leq \nu) &= \frac{P(X = 1)}{P(X = 1) + P(X = 2) + \cdots + P(X = \nu)} \\ &= \frac{C \cdot 1}{C(1 + 2 + \cdots + \nu)} = \frac{2}{\nu(\nu + 1)},\end{aligned}$$

όπου το C είναι η σταθερά που υπολογίσαμε σε προηγούμενο ερώτημα (η οποία απλοποιήθηκε εδώ).

iv) Η μέση τιμή μ της X ισούται με

$$\begin{aligned}\mu = E(X) = \sum_{k=1}^{n} kP(X = k) &= \sum_{k=1}^{n} \frac{2k^2}{n(n+1)} = \frac{2}{n(n+1)} \sum_{k=1}^{n} k^2 \\ &= \frac{2}{n(n+1)} \frac{n(n+1)(2n+1)}{6} = \frac{2n+1}{3}.\end{aligned}$$

Λύση (συνέχεια)

v) Η διακύμανση σ^2 της X ισούται με $\sigma^2 = V(X) = E(X^2) - E(X)^2$. Έχουμε ότι

$$E(X^2) = \sum_{k=1}^{n} k^2 P(X=k) = \frac{2}{n(n+1)} \sum_{k=1}^{n} k^3 = \frac{2}{n(n+1)} \left(\frac{n(n+1)}{2} \right)^2$$
$$= \frac{n(n+1)}{2},$$

οπότε

$$\sigma^2 = E(X^2) - E(X)^2 = \frac{n(n+1)}{2} - \frac{(2n+1)^2}{9}$$
$$= \frac{9n^2 + 9n - 8n^2 + 8n - 2}{18} = \frac{n^2 + n - 2}{18} = \frac{(n-1)(n+2)}{18}.$$

Άσκηση - Τυχαία γραφήματα (Το μοντέλο $G(n,p)$ των Erdos και Rényi)

Κατασκευάζουμε ένα γράφημα δεσμών $G(n,p)$ με n κορυφές, έτσι ώστε για κάθε ζεύγος κορυφών u, v, ο δεσμός $\{u, v\}$ να ανήκει στο γράφημα με πιθανότητα p (ανεξάρτητα από τους άλλους δεσμούς). Συνήθως το p είναι συνάρτηση του n.

- $i)$ Να βρεθεί ο μέσος βαθμός των κορυφών του $G(n,p)$.
- $ii)$ Να βρεθεί ο μέσος αριθμός δεσμών του $G(n,p)$.
- $iii)$ Να βρεθεί ο μέσος αριθμός των τριγώνων του $G(n,p)$.

Λύση

i) Για κάθε κορυφή v, συμβολίζουμε με X_v τον βαθμό της.
Για κάθε κορυφή $u \neq v$, ορίζουμε την ΤΜ

$$Y_{uv} = \begin{cases} 1, & \text{αν υπάρχει ο δεσμός } \{u, v\}, \\ 0, & \text{αν δεν υπάρχει ο δεσμός } \{u, v\}. \end{cases}$$

Προφανώς, ισχύει ότι $X_v = \sum_u Y_{uv}$. Επιπλέον, για κάθε $u \neq v$, ισχύει ότι $E(Y_{uv}) = 1P(Y_{uv} = 1) + 0P(Y_{uv} = 0) = 1p + 0(1-p) = p$.
Άρα, από την γραμμικότητα της μέσης τιμής ισχύει ότι

$$E(X_v) = E(\sum_u Y_{uv}) = \sum_u E(Y_{uv}) = (n-1)p.$$

Επομένως, για να κατασκευάσουμε ένα τυχαίο γράφημα με n κορυφές, με μέσο βαθμό κορυφής c, επιλέγουμε κάθε δεσμό με πιθανότητα $p = c/(n-1)$.

Λύση (συνέχεια)

ii) Έστω Z το πλήθος δεσμών. Γνωρίζουμε ότι $2Z = \sum_v X_v$. Επομένως,

$$E(Z) = E(\frac{1}{2}\sum_v X_v) = \frac{1}{2}\sum_v E(X_v) = \frac{1}{2}\sum_v (n-1)p$$
$$= \frac{n(n-1)}{2}p = \binom{n}{2}p.$$

Επομένως, αν θέλουμε μέσο πλήθος δεσμών $E(Z) = c$, τότε θέτουμε $p = \dfrac{2c}{n(n-1)}$.

Λύση (συνέχεια)

iii) Παρατηρούμε ότι υπάρχουν $\binom{n}{3}$ τριάδες κορυφών, κάποιες από τις οποίες αποτελούν τρίγωνο.
Έστω u, v, w 3 κορυφές του γραφήματος. Η πιθανότητα να αποτελούν τρίγωνο ισούται με $p \cdot p \cdot p = p^3$. Παρατηρήστε ότι η ύπαρξη ή μη ενός τριγώνου επηρεάζει την πιθανότητα ύπαρξης των άλλων τριγώνων που έχουν κοινή μια ή δύο πλευρές με αυτό.
Έστω X ο αριθμός των τριγώνων του γραφήματος.
Για κάθε τριάδα κορυφών u, v, w θεωρούμε την ΤΜ

$$Y_{uvw} = \begin{cases} 1, & \text{αν υπάρχει το τρίγωνο } \{u, v, w\}, \\ 0, & \text{αν δεν υπάρχει το τρίγωνο } \{u, v, w\}. \end{cases}$$

Ισχύει ότι $X = \sum_{u,v,w} Y_{uvw}$. Επιπλέον, για κάθε u, v, w ισχύει ότι

$$E(Y_{uvw}) = 1 \cdot P(Y_{uvw} = 1) + 0 \cdot P(Y_{uvw} = 0) = p^3.$$

Λύση (συνέχεια)

Άρα,

$$E(X) = E(\sum_{u,v,w} Y_{uvw}) = \sum_{u,v,w} E(Y_{uvw}) = \sum_{u,v,w} p^3 = \binom{n}{3} p^3$$
$$= \frac{n(n-1)(n-2)}{6} p^3.$$

Αν επιλέξουμε $p = \frac{c}{n}$, τότε

$$E(X) = \frac{n(n-1)(n-2)}{6} \cdot \frac{c^3}{n^3} \approx \frac{c^3}{6}$$

Επομένως, για να κατασκευάσουμε ένα τυχαίο γράφημα με n κορυφές και k τρίγωνα, θέλουμε $c^3/6 = k$, ή ισοδύναμα $c = \sqrt[3]{6k}$, οπότε επιλέγουμε κάθε δεσμό με πιθανότητα $p = \sqrt[3]{6k}/n$.