Πιθανογεννήτριες - Πολυωνυμική και Υπεργεωμετρική Κατανομή

Μ. Ζαζάνης

Πιθανότητες ΙΙ

23 Μαρτίου 2020

Ορισμός Πιθανογεννήτριας

Έστω X μια διακριτή τυχαία μεταβλητή με κατανομή $\mathbb{P}(X = k) = p_k$, $k = 0, 1, 2, \ldots$. Η πιθανογεννήτρια της X είναι η συνάρτηση

$$G(z) := \sum_{k=0}^{\infty} p_k z^k \qquad (1)$$

Η μεταβλητή z παίρνει πραγματικές τιμές (στις εφαρμογές που θα συζητήσουμε) και κυμαίνεται στο διάστημα στο οποίο η σειρά (1) συγκλίνει. Παρατηρείστε ότι η (1) σημαίνει ότι $G(z) := \mathbb{E}[z^X]$.

Προσδιορισμός Κατανομής από την Πιθανογεννήτρια

Συμβολίζουμε με $G^{(k)}(z)$ την παράγωγο τάξης k υπολογισμένη στην τιμή z. Τότε

$$p_k = \frac{1}{k!} G^{(k)}(0), \qquad k = 0, 1, 2, \ldots, \tag{2}$$

Η $G(z)$ προσδιορίζει με μοναδικά την κατανομή $\{p_k\}$ μέσω της (2).
Αν οι X, Y, είναι *ανεξάρτητες* τυχαίες μεταβλητές με πιθανογεννήτριες $G_X(z)$, $G_Y(z)$ αντίστοιχα, η πιθανογεννήτρια του αθροίσματος $Z = X + Y$ είναι

$$G_Z(z) = G_X(z) G_Y(z).$$

Πράγματι,

$$G_Z(z) = \mathbb{E}[z^{X+Y}] = \mathbb{E}[z^X z^Y] = \mathbb{E}z^X \mathbb{E}z^Y,$$

λόγω της ανεξαρτησίας των X, Y. Επαγωγικά, για οποιοδήποτε πεπερασμένο πλήθος ανεξάρτητων τ.μ. X_i, $i = 1, 2, \ldots, n$ με (κοινή) πιθανογεννήτρια $G_X(z)$, το άθροισμά $S_n := X_1 + \cdots + X_n$ έχει πιθανογεννήτρια $G_{S_n}(z) = (G_X(z))^n$.

Η κατανομή Bernoulli

Η τυχαία μεταβλητή

$$X = \begin{cases} 0 & \text{με πιθανότητα } q \\ 1 & \text{με πιθανότητα } p \end{cases}$$

όπου $p \in [0, 1]$ και $q = 1 - p$ ονομάζεται Bernoulli. Η κατανομή Bernouli ειναι η πλέον στοιχειώδης και αποτελεί δομικό λίθο για πιο περίπλοκες κατανομές. Η πιθανογεννήτρια της κατανομής αυτής είναι

$$G(z) = z^0 \cdot \mathbb{P}(X = 0) + z^1 \cdot \mathbb{P}(X = 1) + z^2 \cdot \mathbb{P}(X = 2) + \cdots \ = \ q + pz.$$

Η μέση τιμή της X είναι $\mathbb{E}[X] = G'(1) = p$. Η δεύτερη παραγοντική ροπή είναι $\mathbb{E}[X(X-1)] = G''(1) = 0$. Συνεπώς $\mathbb{E}[X^2 - X] = 0$ και $\mathbb{E}[X^2] = \mathbb{E}[X]$, δηλαδή $\mathbb{E}[X^2] = p$. Συνεπώς $\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X]) = p - p^2 = p(1-p) = pq$.

Η Διωνυμική κατανομή

Μια τυχαία μεταβλητή X με κατανομή

$$\mathbb{P}(X=k)=\binom{n}{k}p^k q^{n-k}, \qquad \text{αν } k=0,1,2,\ldots,n \text{ και } 0 \text{ διαφορετικά}$$

Πιθανότητα να έχουμε k επιτυχίες σε n ανεξάρτητες δοκιμές.
Πιθανογεννήτρια:

$$G(z)=\sum_{k=0}^{n}\binom{n}{k}p^k q^{n-k} z^k=(q+pz)^n.$$

Ο μέσος και η διασπορά μπορούν να υπολογισθούν από την πιθανογεννήτρια ως εξής: $G'(z)=np(q+pz)^{n-1}$ και $G''(z)=n(n-1)p^2(q+pz)^{n-2}$ (υποθέτουμε ότι $n\geq 2$). $\mathbb{E}[X]=G'(1)=np$ (χρησιμοποιήσαμε το γεγονός ότι $p+q=1$). $\mathbb{E}[X(X-1)]=n(n-1)p^2$. Συνεπώς $\mathbb{E}[X^2]=n(n-1)p^2+\mathbb{E}[X]=n(n-1)p^2+np$ και κατά συνέπεια

$$\begin{aligned}\mathrm{Var}(X) &= \mathbb{E}[X^2]-(\mathbb{E}[X])^2=n(n-1)p^2+np-n^2p^2=np-np^2\\ &= np(1-p)=npq.\end{aligned}$$

Η κατανομή Poisson

Η τυχαία μεταβλητή X έχει κατανομή Poisson με παράμετρο $\alpha > 0$ αν

$$\mathbb{P}(X = k) = \frac{1}{k!}\alpha^k e^{-\alpha}, \qquad k = 0, 1, 2, \dots .$$

Η πιθανογεννήτριά της δίδεται από την

$$G(z) = \sum_{k=0}^{\infty} z^k \frac{1}{k!}\alpha^k e^{-\alpha} = e^{-\alpha} \sum_{k=0}^{\infty} \frac{1}{k!}(\alpha z)^k = e^{-\alpha} e^{z\alpha} = e^{-\alpha(1-z)}.$$

Η μέση τιμή και η διασπορά της κατανομής Poisson υπολογίζεται εύκολα ως $EX = \mathrm{Var}(X) = \alpha$.

Μια από τις σημαντικότερες ιδιότητες της κατανομής Poisson είναι ότι προκύπτει ως το όριο της διωνυμικής κατανομής $\text{Binom}(n, \alpha/n)$ όταν $n \to \infty$ (δηλαδή στην περίπτωση που έχουμε ένα μεγάλο αριθμό ανεξάρτητων δοκιμών, n, η κάθε μια από τις οποίες έχει μικρή πιθανότητα επιτυχίας, α/n). Αυτό είναι εύκολο να διαπιστωθεί εξετάζοντας την πιθανογεννήτρια της διωνυμικής κατανομής $(n, \alpha/n)$ και παίρνοντας το όριο $n \to \infty$. Πράγματι,

$$\lim_{n\to\infty}\left(1-\frac{\alpha}{n}+z\frac{\alpha}{n}\right)^n=\lim_{n\to\infty}\left(1-\frac{\alpha(1-z)}{n}\right)^n=e^{-\alpha(1-z)}$$

το οποίο δείχνει ότι $\text{Binom}(\alpha/n, n) \to \text{Poi}(\alpha)$ όταν $n \to \infty$.
Ισχύει επίσης ότι, αν X_1, X_2 είναι ανεξάρτητες τυχαίες μεταβλητές Poisson με παραμέτρους α_1, α_2 αντίστοιχα, τότε $X_1 + X_2 \sim \text{Poi}(\alpha_1 + \alpha_2)$. Ο απλούστερος τρόπος να το διαπιστώσουμε είναι να εξετάσουμε την πιθανογεννήτρια $Ez^{X_1+X_2} = Ez^{X_1}Ez^{X_2} = e^{-\alpha_1(1-z)}e^{-\alpha_2(1-z)} = e^{-(\alpha_1+\alpha_2)(1-z)}$.

Η Γεωμετρική Κατανομή

Η τυχαία μεταβλητή X είναι γεωμετρική με παράμετρο p όταν η κατανομή της δίδεται από την

$$\mathbb{P}(X = k) = q^{k-1}p, \qquad k = 1, 2, 3, \ldots, \tag{3}$$

όπου $p \in (0, 1)$ και $q = 1 - p$. Η πιθανογεννήτρια συνάρτηση της κατανομής αυτής δίδεται από την σχέση

$$G(z) = \sum_{k=1}^{\infty} q^{k-1}pz^k = \frac{(1-q)z}{1-qz}. \tag{4}$$

Η παράμετρος p ονομάζεται συνήθως ‘πιθανότητα επιτυχίας’ και η τιμή της X είναι ο αριθμός των ανεξάρτητων δοκιμών που απαιτούνται μέχρι την πρώτη επιτυχία αν κάθε δοκιμή έχει πιθανότητα επιτυχίας p. Εναλλακτικά, μπορούμε να εξετάσουμε τον αριθμό των *αποτυχιών*, Y, μέχρι την πρώτη επιτυχία. Στην περίπτωση αυτή $Y = X - 1$ και

$$\mathbb{P}(Y = k) = q^k p, \qquad k = 0, 1, 2, \ldots, \tag{5}$$

με πιθανογεννήτρια

$$\mathbb{E}z^Y = \frac{1-q}{1-qz}. \tag{6}$$

Μπορούμε να δούμε εύκολα ότι $\mathbb{E}Y = q/p$ και $\mathrm{Var}(Y) = q/p^2$. Επίσης, $\mathbb{E}X = 1 + \mathbb{E}Y = 1/p$ και $\mathrm{Var}(X) = \mathrm{Var}(Y) = q/p^2$.

Η Αρνητική Διωνυμική Κατανομή (Κατανομή Pascal)

Ξεκινάμε με τον ορισμό του διωνυμικού συντελεστή στην περίπτωση που $a \in \mathbb{R}$ και $n \in \mathbb{N}$ ως

$$\binom{a}{n} := \frac{a(a-1)\ldots(a-n+1)}{n!}.$$

Αν ο a είναι φυσικός αριθμός τότε $\binom{a}{n} = 0$ για κάθε $n > a$. Αν ο a είναι αρνητικός ακέραιος ή μη ακέραιος πραγματικός αριθμός, τότε $\binom{a}{n} \neq 0$ για κάθε $n \in \mathbb{N}$. Σύμφωνα με το διωνυμικό θεώρημα, για κάθε $|x| < 1$ και $\alpha \in \mathbb{R}$:

$$(1+x)^{\alpha} = \sum_{k=0}^{\infty} \binom{\alpha}{k} x^k. \tag{7}$$

(Αν ο α είναι θετικός ακέραιος τότε $\binom{\alpha}{k} = 0$ για κάθε $k = \alpha+1, \alpha+2, \cdots$ και συνεπώς η άπειρη σειρά (7) γίνεται ένα άθροισμα με πεπερασμένο πλήθος όρων: $(1+x)^{\alpha} = \sum_{k=0}^{\alpha} \binom{\alpha}{k} x^k$.)

Ο διωνυμικός συντελεστής $\binom{-\alpha}{n}$ γράφεται ως

$$\begin{aligned}\binom{-\alpha}{n} &= \frac{(-\alpha)(-\alpha-1)\cdots(-\alpha-n+2)(-\alpha-n+1)}{n!}\\ &= (-1)^n\frac{(\alpha+n-1)(\alpha+n-2)\cdots(\alpha+1)\alpha}{n!}\\ &= (-1)^n\binom{\alpha+n-1}{n}.\end{aligned}$$

Συνεπώς, για κάθε $|x|<1$ ισχύει η ταυτότητα

$$(1-x)^{-\alpha} = \sum_{k=0}^{\infty}\binom{-\alpha}{k}(-x)^k = \sum_{k=0}^{\infty}\binom{\alpha+k-1}{k}x^k. \tag{8}$$

Αν $p \in (0,1)$ και $q = 1 - p$ τότε η *αρνητική διωνυμική κατανομή* με παραμέτρους p και $\alpha > 0$ ορίζεται ως

$$\mathbb{P}(X = k) = \binom{\alpha + k - 1}{k} p^\alpha q^k, \qquad k = 0, 1, 2, \ldots. \tag{9}$$

Η παραπάνω έκφραση ορίζει πράγματι μια κατανομή πιθανότητας στους μη αρνητικούς ακεραίους δεδομένου ότι $\binom{\alpha+k-1}{k} > 0$ όταν $\alpha > 0$ για κάθε $k \in \mathbb{N}$ και $\sum_{k=0}^{\infty} \binom{\alpha+k-1}{k} p^\alpha q^k = p^\alpha (1-q)^{-\alpha} = 1$, λόγω της σχέσης (8).

Η πιθανογεννήτρια της αρνητικής διωνυμικής κατανομής δίδεται από την

$$G(z) = \sum_{k=0}^{\infty} \binom{\alpha+k-1}{k} p^{\alpha} q^k z^k = \left(\frac{p}{1-qz}\right)^{\alpha}.$$

Η μέση τιμή της αρνητικής διωνυμικής τυχαίας μεταβλητής X είναι $\mathbb{E}X = G'(1) = \alpha q \frac{p^{\alpha}}{(1-q)^{\alpha+1}}$ ή

$$\mathbb{E}X = \alpha \frac{q}{p}.$$

Ομοίως, $\mathbb{E}X(X-1) = G''(1) = \alpha(\alpha+1) q^2 \frac{p^{\alpha}}{(1-q)^{\alpha+2}} = \alpha(\alpha+1)\left(\frac{q}{p}\right)^2$. Επομένως έχουμε $\mathbb{E}X^2 = \alpha(\alpha+1)\left(\frac{q}{p}\right)^2 + \alpha\frac{q}{p}$ ανδ τηυς $\mathrm{Var}(X) = \alpha(\alpha+1)\left(\frac{q}{p}\right)^2 + \alpha\frac{q}{p} - \left(\alpha\frac{q}{p}\right)^2 = \alpha\frac{q}{p}\left(1+\frac{q}{p}\right)$ ή

$$\mathrm{Var}(X) = \alpha \frac{q}{p^2}.$$

Όταν $\alpha = m \in \mathbb{N}$ η αρνητική διωνυμική κατανομή είναι η κατανομή του αθροίσματος m ανεξάρτητων τυχαίων μεταβλητών γεωμετρικά κατανεμημένων (του τύπου (5)), η κάθε μια με πιθανότητα επιτυχίας p. Αυτό αποδεικνύεται εύκολα με την βοήθεια των πιθανογεννητριών. Επίσης

$$\mathbb{E}[X(X-1)\cdots(X-k+1)] = G^{(k)}(1). \tag{10}$$

Η έκφραση (10) ονομάζεται κατιούσα παραγοντική ροπή τάξης k. Οι συνήθεις ροπές μπορούν να υπολογιστούν από αυτές.

Πολυωνυμική Κατανομή

Έστω ένα πείραμα που μπορεί να έχει m διαφορετικά αποτελέσματα, $a_1, a_2, \ldots, a_m$. Η πιθανότητα του a_i είναι p_i και $\sum_{i=1}^{m} p_i = 1$. Π.χ. η η ρίψη ενός τίμιου ζαριού όπου υπάρχουν έξι διαφορετικά ισοπίθανα αποτελέσματα.
Εκτελούμε το πείραμα n φορές, ανεξάρτητα. Έστω X_1 ο αριθμός των πειραμάτων με αποτέλεσμα a_1, X_2 ο αριθμός των πειραμάτων με αποτέλεσμα a_2 και X_m ο αριθμός των πειραμάτων με αποτέλεσμα ήταν a_m.
Τα $X_1, X_2, \ldots, X_m$ είναι τυχαίες μεταβλητές οι οποίες δεν είναι ανεξάρτητες μεταξύ τους. Ισχύει ότι $X_1 + X_2 + \cdots + X_m = n$.
Η από κοινού κατανομή των τυχαίων αυτών μεταβλητών ονομάζεται πολυωνυμική. Αν $n_1, n_2, \ldots, n_m$ είναι μη αρνητικοί ακέραιοι, τέτοιοι ώστε $\sum_{i=1}^{m} n_i = n$, τότε

$$P(X_1 = n_1, X_2 = n_2, \ldots, X_m = n_m) = \binom{n}{n_1, n_2, \ldots, n_m} p_1^{n_1} p_2^{n_2} \cdots p_m^{n_m}.$$

Παράδειγμα

Έστω ένα πείραμα που έχει τρία δυνατά αποτελέσματα, τα a_1, a_2, και a_3 τα οποία συμβαίνουν με πιθανότητες p_1, p_2 και p_3, όπου $p_1 + p_2 + p_3 = 1$. Το εκτελούμε 10 φορές ανεξάρτητα. X_i, $i = 1, 2, 3$, είναι ο αριθμός των πειραμάτων των οποίων το αποτέλεσμα ήταν a_i. Ας υπολογίσουμε την πιθανότητα $P(X_1 = 3, X_2 = 5, X_3 = 2)$. Η πιθανότητα να έχουμε τα αποτελέσματα $a_1 a_1 a_1 a_2 a_2 a_2 a_2 a_2 a_3 a_3$ είναι $p_1^3 p_2^5 p_3^2$ λόγω της ανεξαρτησίας των πειραμάτων. Η πιθανότητα μιας οποιασδήποτε άλλης πραγματοποίησης με 3 a_1, 5 a_2 και 2 a_3 είναι η ίδια. *Πρέπει όμως να δούμε πόσες διαφορετικές πραγματοποιήσεις υπάρχουν με τον συγκεκριμένο αριθμό των τριών αποτελεσμάτων, με πόσους διαφορετικούς τρόπους δηλαδή μπορούμε να βάλουμε στη σειρά 3 a_1, 5 a_2 και 2 a_3.* Αυτός είναι ο αριθμός των επαναληπτικών μεταθέσεων που είναι $\binom{10}{3,5,2} = 2520$. Συνεπώς

$$P(X_1 = 3, X_2 = 5, X_3 = 2) = \binom{10}{3, 5, 2} p_1^3 p_2^5 p_3^2.$$

Παράδειγμα 2: Ρίψεις Ζαριού

Η πιθανότητα σε 6 ρίψεις ενός τιμίου ζαριού να έχουμε 2 δυάρια, 2 τεσσάρια και 2 εξάρια είναι

$$\begin{aligned}\binom{6}{0,2,0,2,0,2}\left(\frac{1}{6}\right)^0\left(\frac{1}{6}\right)^2\left(\frac{1}{6}\right)^0\left(\frac{1}{6}\right)^2\left(\frac{1}{6}\right)^0\left(\frac{1}{6}\right)^2 &= \frac{6!}{2!0!2!0!2!0!}\frac{1}{6^6} \\ &= 0,0019.\end{aligned}$$

Η από κοινού κατανομή του αριθμού των διαφορετικών αποτελεσμάτων, $(X_1, X_2, \ldots, X_m)$ ονομάζεται *πολυωνυμική κατανομή*. Το αθροισμα των πολυωνυμικών πιθανοτήτων για όλα τα $n_1, n_2, \ldots, n_m$ τέτοια ώστε $\sum_{i=1}^m n_i = n$ είναι μονάδα λόγω του πολυωνυμικού θεωρήματος:

$$\begin{aligned}\sum_{\{(n_1,\ldots,n_m):\sum_{i=1}^m n_i=n\}}\binom{n}{n_1, n_2, \ldots, n_m} p_1^{n_1} p_2^{n_2}\cdots p_m^{n_m} &= (p_1+\cdots+p_m)^n \\ &= 1^n = 1.\end{aligned}$$

Η περιθώρια κατανομή της X_i μπορεί να υπολογισθεί αθροίζοντας τις πολυωνυμικές πιθανότητες.

Ας δούμε ξανά το πείραμα που έχει τρία διαφορετικά αποτελέσματα, a_1, a_2, και a_3 τα οποία συμβαίνουν με πιθανότητες p_1, p_2 και p_3 αντίστοιχα, όπου $p_1 + p_2 + p_3 = 1$, και το οποίο επαναλαμβάνουμε ανεξάρτητα n φορές. Έστω X_1, X_2, X_3, ο αριθμός των αποτελεσμάτων τύπου a_1, a_2, και a_3 στις n επαναλήψεις. Η από κοινού κατανομή των (X_1, X_2, X_3) θα είναι, για $m = 3$,

$$P(X_1 = i, X_2 = j, X_3 = k) = \binom{n}{i, j, k} p_1^i p_2^j p_3^k. \tag{11}$$

Για να βρούμε την περιθώρια κατανομή του X_1 αρκεί να αθροίσουμε ως εξής:

$$P(X_1 = i) = \sum_{\{(j,k): j+k=n-i,\ j,k \in \mathbb{N}_0\}} \binom{n}{i, j, k} p_1^i p_2^j p_3^k \qquad i = 0, 1, 2, \dots, n.$$

Λαμβάνοντας υπ΄ όψιν ότι $i + j + k = n$ και

$$\binom{n}{i, j, k} = \frac{n!}{(n-i)!i!} \frac{(n-i)!}{(n-i-j)!j!} = \binom{n}{i}\binom{n-i}{j}$$

Η παραπάνω έκφραση γράφεται ισοδύναμα ως

$$\begin{aligned} P(X_1 = i) &= \sum_{j=0}^{n-i} \binom{n}{i} \binom{n-i}{j} p_1^i p_2^j p_3^{n-i-j} = \binom{n}{i} p_1^i \sum_{j=0}^{n-i} \binom{n-i}{j} p_2^j p_3^{n-i-j} \\ &= \binom{n}{i} p_1^i \left(p_2 + p_3\right)^{n-i} = \binom{n}{i} p_1^i \left(1 - p_1\right)^{n-i} \end{aligned} \tag{12}$$

όπου, στην προτελευταία ισότητα χρησιμοποιήσαμε το διωνυμικό θεώρημα και στην τελευταία το γεγονός ότι $p_2 + p_3 = 1 - p_1$. Από την (12) είναι σαφές ότι η X_1 έχει Διωνυμική κατανομή με πιθανότητα επιτυχίας p_1 και αριθμό δοκιμών n. Αυτό μπορεί να το καταλάβει κανείς άμεσα θεωρόντας ότι το πείραμα δεν έχει τρία αποτελέσματα αλλά μόνο δύο: το αποτέλεσμα a_1 που συμβαίνει με πιθανότητα p_1 και το αποτέλεσμα «όχι a_1» (δηλαδή a_2 ή a_3) που συμβαίνει με πιθανότητα $p_2 + p_3 = 1 - p_1$.

Με τον ίδιο ακριβώς τρόπο μπορούμε να αποδείξουμε γενικά (και οι δείκτες δεν πρέπει να μας αποθαρρύνουν) ότι η περιθώριος κατανομή του X_i για την γενική πολυωνυμική κατανομή είναι διωνυμική:

$$P(X_i = n_i) \;=\; \binom{n}{n_i} p_i^{n_i}(1-p_i)^{n-n_i}, \qquad n_i = 0, 1, \ldots, n. \tag{13}$$

Από τις ιδιότητες της διωνυμικής κατανομής συμπεραίνουμε ότι $EX_i = np_i$ και $\mathrm{Var}(X_i) = np_i(1-p_i)$.
Στη συνέχεια μπορούμε να υπολογίσουμε την από κοινού κατανομή των (X_1, X_2) για την πολυωνυμική κατανομή (11) με τρία διαφορετικά αποτελέσματα. Αν i, j είναι δύο μη αρνητικοί ακέραιοι τέτοιοι ώστε $i + j \leq n$ τότε, λαμβάνοντας υπ' όψιν ότι $i + j + k = n$ έχουμε

$$\begin{aligned} P(X_1 = i, X_2 = j) &= \binom{n}{i, j, k} p_1^i p_2^j p_3^k \\ &= \binom{n}{i, j, n-i-j} p_1^i p_2^j (1-p_1-p_2)^{n-i-j}, \qquad i + j \leq n. \end{aligned}$$

Προκειμένου να υπολογίσουμε την συνδιακύμανση των X_1, X_2, υπολογίζουμε πρώτα την μέση τιμή του γινομένου

$$E[X_1X_2] \;=\; \sum_{\{(i,j):\ i+j\leq n,\ i\geq 1,\ j\geq 1\}} ij \binom{n}{i, j, n-i-j} p_1^i p_2^j (1-p_1-p_2)^{n-i-j}$$

Πρόβλημα

Ρίχνουμε ένα ζάρι 10 φορές. Ποιά είναι η πιθανότητα να πάρουμε 4 εξάρια, 3 πεντάρια, 2 τεσσάρια και ένα τριάρι (με οποιαδήποτε σειρά);
(Απ. $\binom{10}{4,3,2,1}\frac{1}{6^{10}}$*.)*

Η πολυμεταβλητή υπεργεωμετρική κατανομή

Ας υποθέσουμε ότι σε μια κάλπη έχουμε K_1 μπάλες χρώματος 1, K_2 μπάλες χρώματος 2, κλπ. και K_m μπάλες χρώματος m. Συνολικά δηλαδή στην κάλπη υπάρχουν $N := K_1 + K_2 + \cdots + K_m$ μπάλες, ασχέτως χρώματος. Από αυτές παίρνουμε, χωρίς επανατοποθέτηση, n μπάλες με $n \leq N$. Έστω X_i ο αριθμός των μπαλών χρώματος i, $i = 1, 2, \ldots, m$ στο δείγμα μεγέθους n που πήραμε. Τότε

$$P(X_1 = n_1, X_2 = n_2, \ldots, X_m = n_m) = \frac{\binom{K_1}{n_1}\binom{K_2}{n_2}\cdots\binom{K_m}{n_m}}{\binom{N}{n}}, \qquad n_i \in \mathbb{N}_0, \quad \sum_{i=1}^{m} n_i = n. \tag{17}$$

Επισημαίνουμε πάλι ότι μπορούμε να χρησιμοποιήσουμε την παραπάνω έκφραση ακόμα και στην περίπτωση που $n_i > K_i$ μια και ο αντίστοιχος διωνυμικός συντελεστής μηδενίζεται. Συνεπώς μπορούμε να χρησιμοποιήσουμε την έκφραση (17) για κάθε διάνυσμα μη αρνητικών ακεραίων $(n_1, \ldots, n_m)$.
Ο λόγος που ισχύει η (17) είναι ότι μπορούμε να διαλέξουμε n_i μπάλες χρώματος i από K_i που υπάρχουν στην κάλπη με $\binom{K}{n_i}$ τρόπους ενώ μπορούμε να διαλέξουμε n μπάλες ασχέτως χρώματος από τις N που υπάρχουν στην κάλπη με $\binom{N}{n}$ τρόπους.

Ας επαληθεύσουμε τώρα ότι

$$\sum_{\{(n_1,\ldots,n_m)\}:\sum_{i=1}^m n_i=n} P(X_1 = n_1, \ldots, X_m = n_m) = 1. \tag{18}$$

Προκειμένου να το δείξουμε αρκεί να δούμε ότι

$$\sum_{\{(n_1,\ldots,n_m)\}:\sum_{i=1}^m n_i=n} \binom{K_1}{n_1}\binom{K_2}{n_2}\cdots\binom{K_m}{n_m} = \binom{N}{n}. \tag{19}$$

Ας εξετάσουμε το γινόμενο

$$(1+x_1)^{K_1}(1+x_2)^{K_2}\cdots(1+x_m)^{K_m}.$$

Ο συντελεστής του όρου $x_1^{n_1}x_2^{n_2}\cdots x_m^{n_m}$ θα είναι

$$\binom{K_1}{n_1}\binom{K_2}{n_2}\cdots\binom{K_m}{n_m}.$$

Αν θέσουμε τώρα $x_1 = x_2 = \cdots = x_m = x$ τότε

$$(1+x)^{K_1}(1+x)^{K_2}\cdots(1+x)^{K_m} = (1+x)^N \tag{20}$$

Με βάση τις παραπάνω παρατηρήσεις, ο συντελεστής του x^n ($n \le N = \sum_{i=1}^m K_i$) στο αριστερό μέλος της (20) δίνεται από το αριστερό μέλος της (19) ενώ ο συντελεστής του x^n στο δεξί μέλος της (20) δίνεται από το δεξί μέλος της (19).

Η περιθώρια κατανομή του X_i είναι εύκολο να διαπιστώσουμε ότι είναι υπεργεωμετρική. Αρκεί να θεωρήσουμε ότι όλες οι μπάλες που δεν είναι χρώματος i είναι κάποιου άλλου χρώματος που δεν μας ενδιαφέρει και να διαπιστώσουμε ότι

$$P(X_i = n_i) = \frac{\binom{K_i}{n_i}\binom{N-K_i}{n-n_i}}{\binom{N}{n}}, \qquad n_i = 0, 1, \ldots, n. \tag{21}$$

Με τον ίδιο τρόπο μπορούμε να δούμε ότι η από κοινού κατανομή των X_i, X_j είναι

$$P(X_i = n_i, X_2 = n_j) = \frac{\binom{K_i}{n_i}\binom{K_j}{n_j}\binom{N-K_i-K_j}{n-n_i-n_j}}{\binom{N}{n}}, \qquad n_i + n_j \leq n. \tag{22}$$

Η μέση τιμή και η διασπορά των X_i, X_j δίνεται από τις γνωστές εκφράσεις για την υπεργεωμετρική κατανομή.

$$EX_i = n\frac{K_i}{N}, \qquad \mathrm{Var}(X_i) = n\,\frac{K_i}{N}\,\frac{N-K_i}{N}\,\frac{N-n}{N-1}. \tag{23}$$

Για να προσδιορίσουμε την συνδιακύμανση αρκεί να δούμε ότι

$$\begin{aligned} EX_1X_2 &= \sum_{\substack{(n_1,n_2):\, n_i\geq 0, n_j\geq 0,\\ \{n_1+n_2\leq n\}}} n_1 n_2 \frac{\binom{K_i}{n_i}\binom{K_j}{n_j}\binom{N-K_i-K_j}{n-n_i-n_j}}{\binom{N}{n}} \\ &= \frac{K_iK_j}{\binom{N}{n}} \sum_{\{(n_1,n_2):\, n_i\geq 1, n_2\geq 1,\; n_1+n_2\leq n\}} \binom{K_i-1}{n_i-1}\binom{K_j-1}{n_j-1}\binom{N-K_i-K_j}{n-n_i-n_j} \end{aligned}$$

Αν θέσουμε $l_i = n_i - 1$, $l_j = n_j - 1$ η παραπάνω σχέση μπορεί να γραφεί ως

$$\begin{aligned} EX_1X_2 &= \frac{K_iK_j}{\binom{N}{n}} \sum_{\left\{\substack{(l_1,l_2):\, l_i\geq 0, l_2\geq 0,\\ l_1+l_2\leq n-2}\right\}} \binom{K_i-1}{l_i}\binom{K_j-1}{l_j}\binom{N-2-(K_i-1)-(K_j-1)}{n-2-l_i-l_j} \\ &= K_1K_2\frac{\binom{N-2}{n-2}}{\binom{N}{n}} = n(n-1)\frac{K_1K_2}{N(N-1)}. \end{aligned}$$

Συνεπώς η συνδιακύμανση των X_1, X_2 είναι

$$\begin{aligned} \mathrm{Cov}(X_1, X_2) &= E[X_1X_2] - EX_1EX_2 = n(n-1)\frac{K_1K_2}{N(N-1)} - n\frac{K_1}{N}\,n\frac{K_2}{N} \\ &= -\frac{n(N-n)}{N-1}\,\frac{K_1}{N}\,\frac{K_2}{N}. \end{aligned} \tag{24}$$

Τέλος ο συντελεστής συσχέτισης προσδιορίζεται εύκολα από την (24) και την (23) και μετά από εύκολες απλοποιήσεις προκύπτει ότι

$$\rho = -\sqrt{\frac{K_iK_j}{(N-K_i)(N-K_j)}}\,. \tag{25}$$

Θεώρημα

Ένα κιβώτιο περιέχει χίλιους λαμπτήρες, 100 από τους οποίους δεν λειτουργούν. Οι χίλιοι αυτοί λαμπτήρες τοποθετούνται τυχαία σε 10 κουτιά. Ποιά είναι η πιθανότητα και οι 100 χαλασμένοι λαμπτήρες να βρεθούν στο ίδιο κουτί; Αν X_i, $i = 1, 2, \ldots, 10$, *είναι ο αριθμός των χαλασμένων δοχείων στο κουτί* i, *ποιά είναι η από κοινού κατανομή των τυχαίων μεταβλητών αυτών;*

Πρόβλημα

Μια κάλπη έχει N_1 *σφαιρίδια που γράφουν τον αριθμό 1,* N_2 *που γράφουν τον αριθμό 2 και* N_3 *που γράφουν τον αριθμό 3. Διαλέγω (χωρίς επανατοποθέτηση)* n *σφαιρίδια. Αν* X_i *ο αριθμός των σφαιριδίων, μεταξύ αυτών που επέλεξα, που γράφουν τον αριθμό* i, $i = 1, 2, 3$, *να βρείτε την από κοινού κατανομή των* X_1, X_2, X_3. *Επίσης να ευρεθούν οι μέσοι, διακυμάνσεις και συνδιακυμάνσεις.*