СЕМ, Л-07
(2020-11-12)

(E) Хипергеометрично разпределение

Имаме N обекта, от които K са маркирани $(0 \leq K \leq N)$. Избират се $n \leq N$ от тях и случайната величина X е броя $k \leq K$ на маркираните обекти. Тогава казваме, че X е разпределено хипергеометрично с параметри N, K и n и отбелязваме с $X \sim \mathrm{Hypergeometric} \equiv \mathrm{Hyp}(N, K, n)$.

Хипергеометричното разпределение описва вероятността да изберем определен брой „специални" обекти при избор без заместване от крайна популация.

(Избор без заместване означава, че когато извадим обект от популацията, не го връщаме обратно преди следващото теглене.)

Класически пример:
В кутия има N топки: K червени и $N - K$ бели.
Изваждаме n топки без връщане. Каква е вероятността да извадим точно k черни топки?

(Това гарантира, че няма да искаме повече специални обекти от наличните или повече неспециални от наличните))

<u>Функция на вероятностите (PMF)</u>:

$$\mathbb{P}(X = k) = \frac{\binom{N}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

Поддръжка:

$$\max(0, n - (N - K)) \leq k \leq \min(n, K)$$

Комбинаторна интерпретация:

- $\binom{N}{k}$ = начини да изберем k специални обекта от K налични
- $\binom{N-K}{n-k}$ = начини да изберем останалите $n - k$ обикновени обекта от $N - K$ налични
- $\binom{N}{n}$ = общ брой начини да изберем n обекта от N

<u>Математическо очакване</u>:

$$\mathbb{E}[X] = n \cdot \frac{K}{N}$$

Доказателство (чрез линейност на очакването и индикатори):
Дефинираме индикаторни случайни величини:

$$X_i = \begin{cases} 1 & \text{ако } i\text{-тият изтеглен обект е специален} \\ 0 & \text{иначе} \end{cases}, i = 1, 2, \dots, n$$

Тогава $X = \Sigma_{i=1}^{n} X_i$.

При избор без заместване, вероятността първият изтеглен обект да е специален е:

$$\mathbb{P}(X_1 = 1) = \frac{N}{K}$$

Но поради симетрия, за всеки i:

$$\mathbb{P}(X_i = 1) = \frac{N}{K}$$

Защо? Можем да си представим, че теглим всички n топки наведнъж, а не последователно. Вероятността дадена топка да е специална е K/N за всяка позиция.

Линейност на очакването:

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i] = \sum_{i=1}^{n} \frac{K}{N} = n \cdot \frac{K}{N}$$

Така:

$$\mathbb{E}[X] = n \cdot \frac{K}{N}$$

Доказателство (чрез директно изчисление със сума):

$$\mathbb{E}[X] = \sum_{k} k \cdot \frac{\binom{N}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

Използваме идентичността: $k\binom{N}{k} = K\binom{K-1}{k-1}$ (за $k \geq 1$)

$$\mathbb{E}[X] = \frac{1}{\binom{N}{n}} \sum_k K \binom{K-1}{k-1} \binom{N-K}{n-k}$$

Нека $j = k - 1$, тогава $k = j + 1, j = 0, 1, \ldots, \ \min(n-1, K-1)$

$$= \frac{K}{\binom{N}{n}} \sum_j \binom{K-1}{j} \binom{N-K}{n-1-j}$$

Използваме тъждество на Вандермонд:

$$\sum_j \binom{K-1}{j} \binom{N-K}{n-1-j} = \binom{N-1}{n-1}$$

Тогава:

$$\mathbb{E}[X] = \frac{K}{\binom{N}{n}} \cdot \binom{N-1}{n-1}$$

Но:

$$\binom{N}{n} = \frac{N}{n} \binom{N-1}{n-1}$$

Следователно:

$$\mathbb{E}[X] = \frac{K}{\frac{N}{n}\binom{N-1}{n-1}} \cdot \binom{N-1}{n-1} = n \cdot \frac{K}{N}.$$

Дисперсия:

$$\mathrm{Var}[X] = n \cdot \frac{K}{N} \cdot \frac{N-K}{N} \cdot \frac{N-n}{N-1}$$

Доказателство (чрез втори момент):

$$\mathbb{E}[X(X-1)] = \sum_k k(k-1) \frac{\binom{N}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

Използваме: $k(k-1)\binom{N}{k} = K(K-1)\binom{K-2}{k-2}$ (за $k \geq 2$)

$$= \frac{K(K-1)}{\binom{N}{n}} \sum_k \binom{K-2}{k-2} \binom{N-K}{n-k}$$

Полагаме $j = k - 2$:

$$= \frac{K(K-1)}{\binom{N}{n}} \sum_j \binom{K-2}{j} \binom{N-K}{n-2-j}$$

По Вандермонд:

$$\sum_j \binom{K-2}{j} \binom{N-K}{n-2-j} = \binom{N-2}{n-2}$$

Така:

$$\mathbb{E}[X(X-1)] = \frac{K(K-1)}{\binom{N}{n}} \cdot \binom{N-2}{n-2}$$

Но:

$$\binom{N}{n} = \frac{N(N-1)}{n(n-1)} \binom{N-2}{n-2}$$

Следователно:

$$\mathbb{E}[X(X-1)] = \frac{K(K-1)}{\frac{N(N-1)}{n(n-1)}} = \frac{n(n-1)K(K-1)}{N(N-1)}.$$

Сега вече може да се върнем към изчисляването на дисперсията:

$$\mathbb{E}[X^2] = \mathbb{E}[X(X-1)] + \mathbb{E}[X] = \frac{n(n-1)K(K-1)}{N(N-1)} + \frac{nK}{N}$$

$$\mathrm{Var}[X] = \mathbb{E}[X^2] - \left(\mathbb{E}[X]\right)^2 = \frac{n(n-1)K(K-1)}{N(N-1)} + \frac{nK}{N} - \frac{n^2K^2}{N^2}$$

$$= \frac{n(n-1)K(K-1)}{N(N-1)} + \frac{nK}{N} - \frac{n^2K^2}{N^2}$$

$$= \frac{n(n-1)K(K-1)N + nKN(N-1) - n^2K^2(N-1)}{N^2(N-1)}$$

$$= \frac{nK[-nN - KN + N^2 + nK]}{N^2(N-1)} = \frac{nK(N-K)(N-n)}{N^2(N-1)}$$

Което е искания резултат.

(Ж) Полиномно разпределение

Полиномно разпределение (Multinomial distribution) е дискретно многовариантно разпределение, което обобщава биномното разпределение за случаи с повече от два възможни резултата върху всеки опит.

Дефиниция и параметри (Полиномно разпределение). Нека имаме експеримент с k възможни резултата (категории). Вероятностите за тези резултати са:

$$p_1, p_2, \ldots, p_k \text{ с } \sum_i^k p_i = 1.$$

Повтаряме експеримента независимо n пъти.
Нека X_i = брой пъти, в които се случва резултат i.

Тогава векторът $\mathbf{X} = (X_1, X_2, \ldots, X_k)$ следва полиномно разпределение с параметри n и $\mathbf{p} = (p_1, \ldots, p_k)$:

$$\mathbb{P}(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k) = \frac{n!}{x_1!x_2!\cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}$$

⊕ Хвърляме зарче 10 пъти. Нека:

- X_1 = брой пъти 1 или 2
- X_2 = брой пъти 3
- X_3 = брой пъти 4, 5 или 6

Тогава $\mathbf{p} = \left(\frac{2}{6}, \frac{1}{6}, \frac{3}{6}\right)$. $n = 10$.
Вероятността да имаме $X_1 = 3, X_2 = 4, X_3 = 2$ е:

$$\mathbb{P}(3,4,2) = \frac{10!}{3!4!2!} \cdot \left(\frac{2}{6}\right)^3 \cdot \left(\frac{1}{6}\right)^4 \cdot \left(\frac{3}{6}\right)^2 = \frac{174}{1944}$$

⊕ Анализ на клиентско поведение (моделиране):

(А) $X_i \sim Ber(p)$: поведението на i-тия клиент в даден магазин (купува или не).

(Б) $X = \Sigma_{i=1}^n X_i$, $X \sim \text{Bin}(n, p)$: броя на извършилите покупка клиенти от първите n клиента.

(В1) $X = \min\left\{k \geq 1 \,\middle|\, \Sigma_{i=1}^k X_i = 1\right\}$, $X \sim \text{Geom}(p)$: броя клиенти които ще платят в брой ДО първия клиент, който ще плати с карта (включително).

(В2) $X = \min\left\{k \geq 1 \,\middle|\, \Sigma_{i=1}^{k} X_i = 1\right\} - 1,\ X \sim \mathrm{Geom}(p):$ броя клиенти които ще платят в брой ПРЕДИ първия клиент, който ще плати с карта.

(Г1) $X = \min\left\{k \geq n \,|\, \Sigma_{i=1}^{n} X_i = r\right\}, X \sim \mathrm{NB}(p)$: броя клиенти, които общо трябва да влязат ДО постигнатето на 5 продажби.

(Г1) $X = \min\left\{k \geq n \,|\, \Sigma_{i=1}^{n} X_i = r\right\} - 1,\ X \sim \mathrm{NB}(p):$ броя клиенти, които общо трябва да влязат ПРЕДИ постигнатето на 5 продажби.

(Д) $X \sim \mathrm{Pois}(\lambda)$, където λ=среден брой клиенти за 1 час: броя клиенти, които влизат в магазина за 1 час.

(Е) Магазинът получава доставка от N еднотипни продукти. От тях $K < N$ са дефектни (производителят е съобщил $(K/N) \times 100\%$ дефектност). Отделът по контрол на качеството произволно избира $n < N$ продукта за тестване без връщане. $X \sim \mathrm{Hyp}(N, K, n)$: брой дефектни продукти в извадка от n.

(Е) Магазинът разполага с три еднакво големи секции:

- Електроника (E)
- Дрехи (C)
- Хранителни стоки (F)

Статистиката показва, че всеки клиент, който влезе, с вероятности:

- $p_E = 0.4$ посещава **само** електроника (и купува или не от там),
- $p_C = 0.35$ посещава **само** дрехите,
- $p_F = 0.25$ посещава **само** хранителните стоки.

(Тук се абстрахираме от клиенти, които посещават повече от една секция, за да имаме чисто полиномно разпределение.)

Нека X_E, X_C, X_F са броят клиенти, посетили съответно електроника, дрехи и хранителни стоки от n на брой независими вличания.

Тогава $(X_E, X_C, X_F) \sim \mathrm{Multinomial}(n;\ p_E = 0.4, p_C = 0.35, p_F = 0.25)$.