СЕМ, Л-06
(2020-11-05)

Коментари:

$$\mathbb{E}\left[\sum_{j=1}^{n} X_j\right] = \sum_{j=1}^{n} \mathbb{E}[X_j]$$, винаги (когато очакванията са добре дефинирани).

$$\mathrm{Var}\left[\sum_{j=1}^{n} X_j\right] = \sum_{j=1}^{n} \mathrm{Var}[X]$$, ако $X_1, X_2, \ldots, X_n$ са независими в съвкупност.

Целочислени случайни величини

Дефиниция (Схема на Бернули).
Схема на Бернули (или последователност на Бернули) е поредица от n независими опити, където всеки опит има само два възможни изхода:

1. **Успех** (Success) с вероятност p, където $0 \le p \le 1$
2. **Неуспех** (Failure) с вероятност $q = 1 - p$

Обикновено успехът се кодира като 1, а неуспехът като 0.

Нека имаме n случайни величини $X_1, X_2, \ldots, X_n$, такива че:

- Всяка X_i има разпределение на Бернули: $X_i \sim \mathrm{Bernoulli}(p)$
- Те са взаимно независими

Тогава векторът $(X_1, X_2, \ldots, X_n)$ описва схема на Бернули с параметри n и p.

(A) Разпределение на Бернули

$X \sim \mathrm{Bernoulli}(p) \equiv \mathrm{Ber}(p)$, ако имаме разпределението:

$$X_i = \begin{cases} 1, & \text{с вероятност } p \text{ (успех)} \\ 0, & \text{с вероятност } q = 1 - p \text{ (неуспех)} \end{cases}$$

(Якоб Бернули (1654–1705), Швейцария)

Математическо очакване (средна стойност):

$$\mathbb{E}[X_i] = p, \mathrm{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = p - p^2 = p(1-p) = pq$$

Дисперсия:

$\mathbb{E}[X^2] = 1^2 \cdot p + 0^2 \cdot (1-p) = p$. Следователно:

$$\mathrm{Var}[X] = \mathbb{E}[X^2] - \big(\mathbb{E}[X]\big)^2 = p - p^2 = pq$$

<u>Пораждаща функция (GPF)</u>:

$$G_X(s) = \mathbb{E}s^X = s^0 q + s^1 p = q + ps = 1 - p + ps$$

(Б) Биномно разпределение

Взимаме само първите n случайни величини от схемата на Бернули: $X_1, X_2, \ldots, X_n$ и образуваме сл. вел. $X = \sum_{i=1}^{n} X_i$, която ни отразява броя успехи измежду n експеримента.

Дефиниция (Биномно разпределение). Случайната величина X има биномно разпределение с параметри n (брой опити) и p (вероятност за успех във всеки опит), означавано като:

$$X \sim \mathrm{Binomial}(n, p) \equiv \mathrm{Bin}(n, p),$$

ако представлява броя успехи в n независими опита на Бернули с вероятност за успех p.

<u>Функция на масата на вероятностите (PMF: Probability Mass Function)</u>:

Комбинаторно избираме k експеримента от общо n, които да са успешни по $\binom{n}{k}$ начина и умножаваме по вероятността за успех p точно k пъти и $n-k$ пъти по вероятността за неуспех $\Rightarrow$

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, k = 0, 1, 2, \ldots, n$$

X	0	1	$\ldots$	k	$\ldots$	n
$\mathbb{P}$	q^n	npq^{n-1}	$\ldots$	$\binom{n}{k} p^k q^{n-k}$	$\ldots$	p^n

k успеха измежду n експеримента всеки един измежду които се случва независимо от останалите с вероятност p за успех

Функция на разпределението (CDF: Cumulative Distribution Function):

$$F_X(k) = \mathbb{P}(X \leq k) = \sum_{i=0}^{\lfloor k \rfloor} \binom{n}{i} p^i (1-p)^{n-i}$$

Математическо очакване:

$$\mathbb{E}[X] = np$$

Доказателство:

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i] = \sum_{i=1}^{n} p = np,$$

където $X_i \sim \mathrm{Bernoulli}(p)$ са независими.

Важно! Тук може да се забележи, че индексът на сумите започва от 1 а не от 0. Защо? Защото представяме X като сума от бернулиеви променливи: $X = X_1 + X_2 + \dots + X_n$. Тоест Тук индексът i минава по ОПИТИТЕ: $1, 2, \dots, n$.

Дисперсия:

$$\mathrm{Var}[X] = np(1-p) = npq$$

Където $q = 1 - p$.

Доказателство:

$$\begin{aligned}
\mathrm{Var}[X] &= V\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathrm{Var}[X_i] && \text{(от независимостта } \perp\!\!\!\perp \text{)} \\
&= \sum_{i=1}^{n} p(1-p) && \text{за } X_i \sim \mathrm{Bernoulli}(p) \\
&= np(1-p) = npq
\end{aligned}$$

Стандартно отклонение:

$$\sigma_X = \sqrt{np(1-p)} = \sqrt{npq}$$

Пораждаща функция (PGF):

$$G_X(x) = \mathbb{E}[s^X] = \sum_{k=1}^{n} s^k \binom{n}{k} p^k (1-p)^{n-k} = (1-p+ps)^n$$

Доказателство:

$$G_X(s) = \mathbb{E}[s^{\sum X_i}] = \prod_{i=1}^{n} \mathbb{E}[s^{X_i}] = \prod_{i=1}^{n} (1-p+ps) = (1-p+ps)^n.$$

Имайки пораждащата функция може да сведем проблема за извеждането на функцията на масата на вероятностите, математическото очакване и дисперсията до чисто аналитичен:

$$\mathbb{P}(X=k)=\frac{G_X^{(k)}(0)}{k!}=\frac{\left((1-p+ps)^n\right)^{(k)}\Big|_{s=0}}{k!}$$
$$=\frac{\left(\sum_{k=0}^{n}\binom{n}{k}(ps)^k(1-p)^{n-k}\right)^{(k)}\Big|_{s=0}}{k!}=\binom{n}{k}p^k(1-p)^{n-k}$$

$$G_X(s)'=((1-p+ps)^n)'=n(1-p+ps)^{n-1}(1-p+ps)'$$
$$=np(1-p+ps)^{n-1}.$$

$$G_X(s)''=(np(1-p+ps)^{n-1})'=np(n-1)(1-p+ps)^{n-2}(1-p+ps)'$$
$$=n(n-1)p^2s(1-p+ps)^{n-2}.$$

$$\mathbb{E}[X]=G_X'(1)=np(1-p+ps)^{n-1}\Big|_{s=1}=np.$$

$$\mathrm{Var}[X]=G_X''(1)+G_X'(1)-\left[G_X'(1)\right]^2=n(n-1)p^2+np-n^2p^2=np(1-p).$$

⊕ Колко заразени биха влезли в България от чужбина, ако вероятността човек да е заразен е $p=1\,\%$ и са се върнали $n=100{,}000$ души. Тогава, не лош модел би бил следната случайна величина $X\sim\mathrm{Bin}(n,p)$, която брой колко от върналите се в България са заразени. В случая неинтуитивно приемаме за успех вероятността човек да е заразен.

(В) Геометрично разпределение

При геометричното разпределение може да си мислим за последния пример, но по следния начин: Колко незаразени хора ще влязат в България, преди да влезе първия заразен, върнал се от чужбина?

Геометричното разпределение моделира броя на неуспехите преди първия успех (или броя на опитите до и включително първия успех) в поредица от независими опити на Бернули.

Има два основни варианта:

Вариант 1: Брой опити ДО първия успех (включително успеха)

$$X\sim\mathrm{Geom}(p),\ \mathbb{P}(X=k)=(1-p)^{k-1}p,\ k=1,2,3,\dots$$

Тук X е броят опити до и включително първия успех.

Вариант 2: Брой неуспехи ПРЕДИ първия успех

$$Y \sim \mathrm{Geom}(p), \mathbb{P}(Y = k) = (1-p)^k p, k = 0, 1, 2, \ldots$$

Тук Y е броят неуспехи преди първия успех.

Връзка: $Y = X - 1$, тоест $\mathbb{P}(Y = k) = \mathbb{P}(X = k + 1)$

Дефиниция (Геометрично разпределение).
$X \in \mathrm{Geom}(p)$ и $X = \min\left\{k \geq 1 \,\middle|\, \Sigma_{i=1}^{k} X_i = 1\right\}$. Тоест най-малкото k, за което сумата $\Sigma_{i=1}^{k} X_i$ става единица. Тъй като от тази сума НЕ вадим 1-ца, за да извадим последната стъпка (тази която носи събитието дефинирано като успех), то това е броя неуспехи до първи успех, включително успеха, тоест вариант 1.

$$\oplus\ \underbrace{0, 0, 0,}_{\text{неуспехи}}\ \underbrace{1}_{\text{успех}} \Rightarrow X = 3; \oplus\ \underbrace{0,}_{\text{неуспех}}\ \underbrace{1}_{\text{успех}} \Rightarrow X = 1; \oplus\ \underbrace{1}_{\text{успех}} \Rightarrow X = 0$$

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

$$\mathbb{P}(X = k) = (1-p)^k p, k = 0, 1, 2, \ldots$$

X	0	1	2	$\ldots$	k	$\ldots$
$\mathbb{P}$	p	pq	pq^2	$\ldots$	pq^k	$\ldots$

k неуспеха преди първия успех

<u>Функция на разпределението (CDF)</u>:

$$F_X(m) = \mathbb{P}(X \leq m) = 1 - (1-p)^m$$

Доказателство:

$$\mathbb{P}(X \leq m) = \sum_{k=1}^{m} (1-p)^{k-1} p$$

Това е крайна геометрична прогресия с:
- Първи член $a = p$
- Частно $r = 1 - p$
- Брой членове m

Сумираме:

$$\begin{aligned} F_X(m) &= p \cdot \frac{1-(1-p)^m}{1-(1-p)} \\ &= p \cdot \frac{1-(1-p)^m}{p} = 1 - (1-p)^m, \qquad m = 1, 2, 3, \ldots \end{aligned}$$

За $m < 1$ очевидно $F_X(m) = 0$.

<u>Математическо очакване</u>:

$$\mathbb{E}[X] = \frac{1}{p}$$

Доказателство 1 (чрез сума):

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} k(1-p)^{k-1}p$$

Използваме познат ред: $\Sigma_{k=1}^{\infty} k r^{k-1} = \dfrac{1}{(1-r)^2}$ за $|r| < 1$ (доказва се като вземем първата производна на геометричната прогресия $\Sigma_{k=0}^{\infty} r^k = \dfrac{1}{1-r}$ за $|r| < 1$).

Тук $r = 1 - p$, следователно

$$\mathbb{E}[X] = p \cdot \frac{1}{(1-(1-p))^2} = p \cdot \frac{1}{p^2} = \frac{1}{p}$$

Доказателство 2 (чрез безпаметност):

Нека $\mu = \mathbb{E}[X]$. При първи опит:

- С вероястност p имаме успех и броим 1 опит.
- С вероятност $1 - p$ имаме неуспех, изхарчили сме 1 опит, и поради безпаметността (паметта се нулира), очакваният брой допълнителни опити до успех отново е μ.

Следователно имаме уравнението:

$$\mu = 1 \cdot p + (1 + \mu) \cdot (1 - p)$$
$$\mu = p + (1-p) + (1-p)\mu = 1 + (1-p)\mu$$
$$\mu - (1-p)\mu = 1 \Rightarrow p\mu = 1 \Rightarrow \mu = \frac{1}{p}$$

<u>Дисперсия</u>:

$$\mathrm{Var}[X] = \frac{1-p}{p^2} = \frac{q}{p^2}$$

Доказателство (чрез момента от втори ред):

$$\mathbb{E}[X^2] = \sum_{k=1}^{\infty} k^2(1-p)^{k-1}p$$

Използваме познат ред: $\Sigma_{k=1}^{\infty} k^2 r^{k-1} = \dfrac{1+r}{(1-r)^3}$ *(доказва се като вземем първата производна на геометричната прогресия* $\Sigma_{k=0}^{\infty} r^k = \dfrac{1}{1-r}$ *за* $|r| < 1$*).*

Тук $r = 1 - p$:

$$\mathbb{E}[X^2] = p \cdot \frac{1+(1-p)}{(1-(1-p))^3} = p \cdot \frac{2-p}{p^3} = \frac{2-p}{p^2}.$$

$$\operatorname{Var}[X] = \mathbb{E}[X^2] - \left(\mathbb{E}[X]\right)^2 = \frac{2-p}{p^2} - \frac{1}{p^2} = \frac{1-p}{p^2} = \frac{q}{p^2}.$$

<u>Стандартно отклонение</u>:

$$\sigma_X = \sqrt{\operatorname{Var}[X]} = \frac{q}{\sqrt{p}}$$

<u>Пораждаща функция (PGF)</u>:

$$G_X(s) = \frac{ps}{1-s(1-p)} = \frac{ps}{1-qs},\ |s| < \frac{1}{1-p}$$

Доказателство:

$$G_X(x) = \mathbb{E}[s^X] = \sum_{k=1}^{\infty} s^k \cdot \mathbb{P}(X=k)$$
$$= \sum_{k=1}^{\infty} s^k \cdot q^{k-1} p = ps \sum_{k=1}^{\infty} (sq)^{k-1}.$$

Забелязваме, че това е геометрична прогресия:

$$\sum_{k=1}^{\infty} (sq)^{k-1} = 1 + (sq) + (sq)^2 + \cdots$$

Този ред е сходящ тогава и само тогава, когато $|sq| < 1$, тоест $|s| < \dfrac{1}{q}$.

От формулата за сума на безкрайна геометрична прогресия:

$$\sum_{k=1}^{\infty} (sq)^{k-1} = \frac{1}{1-sq},\ |sq| < 1$$

Следователно:

$$G_X(s) = ps \cdot \frac{1}{q-sq} = \frac{ps}{1-s(1-p)},\ |s| < \frac{1}{1-p}$$

Отново, имайки пораждащата функция може да сведем проблема за извеждането на математическото очакване и дисперсията до чисто аналитичен (само за тях, тъй като функцията $G_X(s)$ е сложна за намирането на k -та производна за функцията на масата на вероятностите):

$$\begin{aligned}
G_X(s)' &= \left(\frac{ps}{1-(1-p)s}\right)' = \frac{(ps)'(1-(1-p)s) - ps(1-(1-p)s)'}{(1-(1-p)s)^2} \\
&= \frac{p(1-s+ps) - ps(-1+p)}{(1-(1-p)s)^2} = \frac{p - \cancel{ps} + \cancel{p^2 s} + \cancel{ps} - \cancel{p^2 s}}{(1-(1-p)s)^2} \\
&= \frac{p}{(1-(1-p)s)^2}\,.
\end{aligned}$$

$$\begin{aligned}
G_X(s)'' &= \left(\frac{p}{(1-(1-p)s)^2}\right)' = (p(1-(1-p)s)^{-2})' \\
&= -2p(1-(1-p)s)^{-3}(1-(1-p)s)' = \\
&= \frac{2p(1-p)}{(1-(1-p)s)^3}\,.
\end{aligned}$$

$$\mathbb{E}[X] = G'_X(1) = \frac{p}{p^2} = \frac{1}{p}.$$

$$\begin{aligned}
\mathrm{Var}[X] &= G''_X(1) + G'_X(1) - \left[G'_X(1)\right]^2 = \frac{2p(1-p)}{p^3} + \frac{1}{p} - \frac{1}{p^2} = \\
&= \frac{2p - 2p^2 + p^2 - p}{p^3} = \frac{p(1-p)}{p^3} = \frac{1-p}{p^2} = \frac{q}{p^2}\,.
\end{aligned}$$

По аналогичен начин може да докажем за Y:

$G_Y(s) = \mathbb{E}[s^Y] = \dfrac{p}{1-(1-p)s}$, $|s| < \dfrac{1}{1-p}$.

$F_Y(m) = \mathbb{P}(Y \leq m) = 1 - (1-p)^{m+1}, n = 0, 1, 2, \ldots.$

$\mathbb{E}[Y] = \dfrac{1-p}{p} = \dfrac{q}{p}$.

$\mathrm{Var}[Y] = \mathrm{Var}[X]$ и съответно $\sigma_Y = \sigma_X$.

Връзка: $Y = X - 1$.

Внимание: Разликата е важна! Проверявай коя дефиниция се използва.

Свойство БЕЗПАМЕТНОСТ (Memoryless)

За всяко $m, n \geq 0$:

$$\mathbb{P}(X > m + n \,|\, X > m) = \mathbb{P}(X > n)$$

Интерпретация: Ако вече сме имали m неуспеха, разпределението на оставащите опити до успех е същото като в началото. Системата „не помни“ миналите неуспехи.

Доказателство:

$$\begin{aligned}\mathbb{P}(X > m + n \,|\, X > m) &= \frac{\mathbb{P}(X > m + n \textbf{ и } X > m)}{\mathbb{P}(X > m)} \\ &= \frac{\mathbb{P}(X > m + n)}{\mathbb{P}(X > m)} = \frac{1 - F_X(m + n)}{1 - F_X(m)} \\ &= \frac{(1 - p)^{m+n}}{(1 - p)^m} = (1 - p)^n = \mathbb{P}(X > n)\,.\end{aligned}$$

Геометричното е единственото дискретно разпределение с това свойство!

⊕ (Вариант 1) Брой хвърляне на монета до първо „ези“: $X \sim \text{Geom}\big(p = 1/2\big)$.

⊕ (Вариант 2) Брой „тури“ при хвърляне на монета до първо „ези“: $Y \sim \text{Geom}\big(p = 1/2\big)$.

⊕ (Вариант 1) Брой хвърляния на зар до първата шестица: $X \sim \text{Geom}(p = 1/6)$.

⊕ (Вариант 2) Брой неуспешни атаки в баскетболен мач до първи кош. Ако грубо казано не се променят играчите и имат вероятност за успех, например $80\,\%$ за кош при атака, то броя на неуспешните атаки до първия успех може да се моделира добре като се допусне, че е разпределен геометрично с параметър $p = 0.8$. Тоест $Y \sim \text{Geom}(p = 4/5)$

(Г) Отрицателно биномно разпределение

Отрицателното биномно разпределение описва броя на опитите на Бернули, необходими за постигане на r успеха (където r е фиксирано цяло число).

Интуиция:
- При геометричното разпределение чакаме 1 успех
- При отрицателно биномното разпределение чакаме r успеха (обобщение)

Както при геометричното, така и тук имаме **два основни варианта**:

<u>Вариант 1: Брой опити ДО r-тия успех (включително успеха)</u>

$$X \sim \text{NBin}(p) \equiv \text{NB}(p),\, \mathbb{P}(X = k) = \binom{k-1}{r-1} p^r (1 - p)^{k-r},\, k = r,\, r + 1,\, r + 2,\, \dots$$

Тук X е броят опити ДО първия успех (включително).

Интерпретация:
- Трябва точно r успеха и $k - r$ неуспеха

- Последният опит (k-тият) трябва да е успех
- Първите $k-1$ опита трябва да имат точно $r-1$ успеха
- Брой начини да подредим тези $r-1$ успеха в $k-1$ опита: $\binom{k-1}{r-1}$

<u>Вариант 2: Брой неуспехи ПРЕДИ r-тия успех</u>

$$
\begin{aligned}
Y \sim \mathrm{NB}(p),\ \mathbb{P}(Y=k) &= \binom{k+r-1}{r-1} p^r (1-p)^k \\
&= \binom{n+r-1}{n} p^r (1-p)^n, \qquad k = 0, 1, 2, \dots
\end{aligned}
$$

Тук Y е броят неуспехи преди първия успех.

Връзка: $Y = X - r$, тоест $\mathbb{P}(Y=k) = \mathbb{P}(X = k+r)$
Специални случаи

Отрицателно биномно разпределение → геометрично разпределение:
Когато $r = 1$:

$$
\mathbb{P}(X=k) = \binom{k-1}{0} p^1 (1-p)^{k-1} = p(1-p)^{k-1}
$$

Това е точно геометрично разпределение (вид 1).

Отрицателно биномно разпределение → биномно разпределение:
За биномното: иумаме фиксиран брой опити n, броим успехите.
Тук имаме фиксиран брой успехи r, броим опитите.
Това са „двойноцо“ - едното е „огледалният образ“ на другото.

Дефиниция (Отрицателно биномно разпределение).
Отново ще разгледаме вариант 1:

$$
X = \min\left\{k \geq 1 \,\middle|\, \sum_{i=1}^{k} X_i = r\right\}.
$$

Тоест най-малкото k, за което сумата $\Sigma_{i=1}^{k} X_i$ става r. Тъй като от тази сума НЕ вадим 1-ца, за да извадим последната стъпка (тази която носи r-тото събитие дефинирано като успех), то това е броя на опитите до r-тия успех (включително).

$\oplus$ $\underbrace{0, 0,}_{\text{неуспехи}}\ \underbrace{1, 1,}_{\text{успехи}}\ \underbrace{0, 0, 0, 0,}_{\text{неуспехи}}\ \underbrace{1}_{\text{успех}} \Rightarrow X = 9, r = 3$

$\oplus$ $\underbrace{1, 1, 1, 1}_{\text{успехи}} \Rightarrow X = 4, r = 4$

Функция на вероятностите (PMF):

$$\mathbb{P}(X=k)=\binom{k-1}{r-1}p^r(1-p)^{k-r},\, k=r,\, r+1,\, r+2,\, \dots$$

X	0	1	$\dots$	k	$\dots$
$\mathbb{P}$	0	p	$\dots$	$\binom{k-1}{r-1}p^r(1-p)^{k-r}$	$\dots$

k опита до
r-тия успех (вкл.)

Функция на разпределението (CDF):

$$F_X(k)=\mathbb{P}(X\le k)=\sum_{i=r}^{k}\mathbb{P}(X=i)=\sum_{i=r}^{k}\binom{i-1}{r-1}p^r(1-p)^{i-r},$$

$k\ge r$ (за $k<r$ вероятността е 0).

Ключова връзка с биномното разпределение:

$$F_X(k)=1-\sum_{i=0}^{r-1}\binom{k}{i}p^i(1-p)^{k-i}.$$

Доказателство:

Събитието $\{X\le k\}$ означава, че сме постигнали r-тия успех най-късно на k-тия опит. Това е еквивалентно на: в първите k опита имаме поне r успеха.
Neka $B\sim \mathrm{Binomial}(k,p)$ е броят успехи в k опита.

Тогава:

$$\begin{aligned}\mathbb{P}(X\le k)&=\mathbb{P}(B\ge r)=1-\mathbb{P}(B\le r-1)\\&=1-\sum_{i=0}^{r-1}\binom{k}{i}p^i(1-p)^{k-i}.\end{aligned}$$

Пораждаща функция (PGF):

$$G_X(s)=\left(\frac{ps}{1-s(1-p)}\right)^r=\left(\frac{ps}{1-qs}\right)^r,\ |s|<\frac{1}{1-p}$$

Доказателство:

$$\begin{aligned}G_X(x)=\mathbb{E}[s^X]&=\sum_{k=1}^{\infty}s^k\cdot\mathbb{P}(X=k)\\&=p^r\sum_{k=r}^{\infty}\binom{k-1}{r-1}(qs)^{k-r}\cdot s^r\end{aligned}$$

Правим смяна на индекса. Нека $j = k - r$, тогава $k = j + r$ и $j = 0, 1, 2, \ldots$

$$G_X(s) = (ps)^r \sum_{j=0}^{\infty} \binom{j+r-1}{r-1} (qs)^j$$

Използваме биномен ред. Знаем, че за $|x| < 1$:

$$(1-x)^{-r} = \sum_{j=0}^{\infty} \binom{j+r-1}{r-1} x^j$$

Това е отрицателният биномен ред.
В нашия случай $x = qs$, така че:

$$\sum_{j=0}^{\infty} \binom{j+r-1}{r-1} (qs)^j = (1-qs)^{-r}, \ |qs| < 1$$

Краен резултат

$$G_X(s) = (ps)^r \cdot (1-qs)^{-r} = \left(\frac{ps}{1-qs}\right)^r$$

Теорема (Отрицателното биномно разпределение като независими геометрични разпределени случайни величини).

Ако $X_1, X_2, \ldots, X_r$ са независими геометрично разпределени случайни величини с параметър p, то:

$$X_1 + X_2 + \cdots X_r \sim \mathrm{NB}(r, p)$$

Интерпретация:

- $X_1 =$ брой опити до 1-ви успех
- $X_2 =$ брой опити до 2-ри успех (след първия)
- $\ldots$
- $X_r =$ брой опити до r-ти успех (след $r-1$-вия)
- Общо опити до r-тия успех: $X_1, X_2, \ldots, X_r$

Доказателство чрез PGF:

PGF на геометрично (брой опити до успех):

$$G_{geom}(s) = \frac{ps}{1-(1-p)s}$$

PGF на отрицателно биномно (брой опити до r успеха):

$$G_{nb}(s) = \left(\frac{ps}{1 - (1 - p)s} \right)^r$$

Но това е точно r-та степен на PGF на геометричното!

$$G_{nb} = [G_{geom}(s)]^r$$

А PGF на сума от независими случайни величини е произведение на техе PGF-и.

Въпреки тази връзка, отрицателното биномно разпределение НЯМА свойството безпаметност (при $r > 1$), тъй като ако вече имаме някакъв брой успехи, това променя разпрелеадението на оставащите опити.
<u>Математическо очакване</u>:

$$\mathbb{E}[X] = \frac{r}{p}$$

Доказателство 1 (чрез сума):

$$\mathbb{E}[X] = \sum_{k=r}^{\infty} k \cdot \binom{k-1}{r-1} p^r q^{k-r}$$

Използваме идентичността: $k \binom{k-1}{r-1} = r \binom{k}{r}$.

Доказателство на идентичността:

$$k \binom{k-1}{r-1} = k \cdot \frac{(k-1)!}{(r-1)!(k-r)!} = \frac{k!}{(r-1)!(k-r)!}$$

$$r \binom{k}{r} = r \cdot \frac{k!}{r!(k-r)!} = \frac{k!}{(r-1)!(k-r)!}$$

Тогава,

$$\mathbb{E}[X] = \sum_{k=r}^{\infty} r \binom{k}{r} p^r q^{k-r} = rp^2 \sum_{k=r}^{\infty} q^{k-r}.$$

Полагаме $m = k - r$ (тогава $k = m + r, m = 0, 1, 2, \ldots$):

$$= rp^r \sum_{m=0}^{\infty} \binom{m+r}{r} q^m.$$

Използваме отрицателен биномен ред:

$$\sum_{m=0}^{\infty} \binom{m+r}{r} q^m = (1-q)^{-(r+1)}, \text{ за } |q| < 1$$

Но $(1-q) = p,$ така че $(1-q)^{-(r+1)} = p^{-(r+1)}$. Тогава:

$$\mathbb{E}[X] = rp^r \cdot p^{-(r+1)} = rp^{-1} = \frac{r}{p}$$

Доказателство 2 (чрез сума от геометрични):

Знаем, че ако $X \sim \mathrm{NB}(r, p)$, то X може да се представи като:

$$X = X_1 + X_2 + \cdots + X_r$$

Където $X_i \sim \mathrm{Geom}(p)$ са независими и всеки X_i е брой до първи успех.
За геометрично разпределение знаем:

$$\mathbb{E}[X_i] = \frac{1}{p}$$

По линейност на очакването:

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^{r} X_i\right] = \sum_{i=1}^{r} \mathbb{E}[X_i] = r \cdot \frac{1}{p} = \frac{r}{p}.$$

Доказателство 3 (чрез PGF):

PGF на X е:

$$G_X(s) = \left(\frac{ps}{1-qs}\right)^r$$

$$\begin{aligned}
G_X'(s) &= r\left(\frac{ps}{1-qs}\right)^{r-1} \cdot \frac{\mathrm{d}}{\mathrm{d}s}\left(\frac{ps}{1-qs}\right) \\
&= r\left(\frac{ps}{1-qs}\right)^{r-1} \cdot \frac{p(1-qs) - ps(-q)}{(1-qs)^2} = \\
&= r\left(\frac{ps}{1-qs}\right)^{r-1} \cdot \frac{p - \cancel{pqs} + \cancel{pqs}}{(1-qs)^2} = r\left(\frac{ps}{1-qs}\right)^{r-1} \cdot \frac{p}{(1-qs)^2}
\end{aligned}$$

При $s = 1$:

$$G_X'(1) = r\left(\frac{p}{1-q}\right) \cdot \left(\frac{p}{(1-q)^2}\right) = r\left(\frac{p}{p}\right)^{r-1} \cdot \frac{p}{p^2} = r \cdot 1 \cdot \frac{1}{p} = \frac{r}{p}$$

Аналогично се доказва, че за Вариант 2: $\mathbb{E}[X] = \dfrac{r(1-p)}{p} = \dfrac{rq}{p}$.

<u>Дисперсия</u>:

$$\mathrm{Var}[X] = \frac{1-p}{p^2} = \frac{q}{p^2}$$

Доказателтво:

Ще го направим само по елегантния начин чрез сума от геометрично разпределени случайни величини.

$$\mathrm{Var}[X] = \mathrm{Var}\left[\sum_{i=1}^{r} Y_i\right] \underbrace{=}_{\text{незав. } \perp\!\!\!\perp} \sum_{i=1}^{r} \mathrm{Var}[Y_i] = r \cdot \frac{q}{p^2} = \frac{rq}{p^2} = \frac{r(1-p)}{p^2}.$$

<u>Стандартно отклонение</u>:

$$\sigma_X = \sqrt{\mathrm{Var}[X]} = \frac{\sqrt{r(1-p)}}{p} = \frac{\sqrt{rq}}{p}.$$

Аналичичен подход за извеждане на PMF чрез PGF:

Знаем, че:

$$G_X(s) = \left(\frac{ps}{1-qs}\right)^r, q = 1-p, |s| < \frac{1}{q}$$

където $X = r, r+1, r+2, \ldots$

Развиваме $G_X(s)$ в ред на Тейлър около $s = 0$.
Първо развиваме $(1-qs)^{-r}$ с отрицателен биномен ред:

$$(1-qs)^{-r} = \sum_{n=0}^{\infty} \binom{n+r-1}{r-1} (qs)^n$$

за $|qs| < 1$.
Тогава:

$$G_X(s) = p^r s^r \sum_{n=0}^{\infty} \binom{n+r-1}{r-1} q^n s^n = p^r \sum_{n=0}^{\infty} \binom{n+r-1}{r-1} q^n s^{n+r}$$

Правим смяна на индекса $k = n + r$, тогава $n = k - r$ и $k = r, r+1, r+2, \ldots$

$$
\begin{aligned}
G_X(s) &= \sum_{k=r}^{\infty} p^r \binom{(k-r)+r-1}{r-1} q^{k-r} \\
&= \sum_{k=r}^{\infty} p^r \binom{k-1}{r-1} q^{k-r} s^k .
\end{aligned}
$$

Тъй като $G_X(s) = \Sigma_{k=r}^{\infty}\mathbb{P}(X=k)s^k$, сравняваме коефициентите:

$$
\mathbb{P}(X=k) = p^r \binom{k-1}{r-1} q^{k-r},\ k = r, r+1, r+2, \dots
$$

(Д) Поасоново разпределение

Поасоновото разпределение моделира броя на събития, случващи се в фиксиран интервал от време или пространство, при следните условия:

- Събитията се случват независимо
- Средният брой събития в интервала е постоянен (обозначава се с λ)
- Две събития не могат да се случват точно в един и същи момент
- Вероятността за събитие в много малък интервал е пропорционална на дължината на интервала

Дефиниция (Поасоново разпределение). Нека $\lambda > 0$. Казваме, че X е поасоново разпределена случайна величина с параметър λ и бележим с $X \sim \text{Pois}(\lambda)$, ако:

$$
\mathbb{P}(X=k) = \frac{\lambda^k}{k!}e^{-\lambda},\ k \geq 0
$$

Проверка (за това дали функцията може да е вероятност):

$$
1 \stackrel{?}{=} \sum_{k=0}^{\infty} \mathbb{P}(X=k) = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!}e^{-\lambda} = e^{-\lambda} \cdot \underbrace{\sum_{k=0}^{\infty} \frac{\lambda^k}{k!}}_{\substack{\text{развитие в ред} \\ \text{на Тейлър за } e^{\lambda}}} = e^{-\lambda} \cdot e^{\lambda} = 1
$$

Какво може да се моделира добре с поасоново разпределена случайна величина:

⊕ Брой постъпили заявки към сървър за единица време
⊕ Брой катастрофи за единица време на дадено кръстовище или територия
⊕ Брой насекоми за единица площ
⊕ Брой голове за единица време
⊕ Брой звезди в някакъв регион
⊕ Брой получени обаждания в кол център на час

Пораждаща функция (PGF):

$$G_X(s) = \mathbb{E}[s^X] = \sum_{k=0}^{\infty} s^k \cdot \mathbb{P}(X=k) = \sum_{k=0}^{\infty} s^k \cdot \frac{\lambda^k e^{-\lambda}}{k!}$$
$$= e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda s)^k}{k!} = e^{-\lambda} \cdot e^{\lambda s} = e^{\lambda(s-1)}.$$

$$G'_X(x) = \left(e^{\lambda(s-1)}\right)' = \left(\lambda(s-1)\right)' e^{\lambda(s-1)} = \lambda e^{\lambda(s-1)}$$

$$G''_X(x) = \left(\lambda e^{\lambda(s-1)}\right)' = \left(\lambda(s-1)\right)' \lambda e^{\lambda(s-1)} = \lambda^2 e^{\lambda(s-1)}$$

Математическо очакване:

$$\mathbb{E}[X] = \lambda$$

Доказателство:

$$\mathbb{E}[X] = G'_X(1) = \lambda e^{\lambda(1-1)} = \lambda e^0 = \lambda\,.$$

Дисперсия:

$$\mathrm{Var}[X] = \lambda$$

Доказателство:

$$\mathrm{Var}[X] = G''_X(1) + G'_X(1) - \left[G_x(1)\right]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

Теорема (Симеон Дени Поасон (1781–1840), Франция).

Нека $X_n \sim \mathrm{Bin}(n, p_n)$, където вероятността за успех p_n зависи от n.
Ако $n \to \infty$ и $p_n \to 0$ така, че $np_n \to \lambda > 0$, тогава за всяко фиксирано $k = 0, 1, 2, \ldots$:

$$\lim_{n\to\infty} \mathbb{P}(X_n = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Тоест, биномното разпределение се доближава до поасоново с параметър λ.

Интуиция
Когато имаме:

- Много опити (n е голямо)
- Много рядък успех (p е много малко)
- Но средният брой успехи np остава постоянен ($\approx \lambda$)

Тогава биномното разпределение става неразличимо от поасоновото

Това напрактика означава, че с поасоновото разпределение може много добре да приближаваме биномното разпределение, когато имаме много на брой опити.

Доказателство:

За $X_n \sim \text{Bin}(n, p_n)$:

$$\mathbb{P}(X_n = k) = \binom{n}{k} p_n^k (1 - p_n)^{n-k} = \frac{n(n-1)\cdots(n-k+1)}{k!} \cdot p_n^k \cdot (1 - p_n)^{n-k}$$

Нека $\lambda_n = np_n$. По условие $\lambda_n \to \lambda$ при $n \to \infty$. Тогава $p_n = \dfrac{\lambda_n}{n}$. Заместваме:

$$\begin{aligned}
\mathbb{P}(X_n = k) &= \frac{n(n-1)\cdots(n-k+1)}{k!} \cdot \left(\frac{\lambda_n}{n}\right)^k \cdot \left(1 - \frac{\lambda_n}{n}\right)^{n-k} \\
&= \frac{n(n-1)\cdots(n-k+1)}{n^k} \cdot \frac{\lambda_n^k}{k!} \cdot \left(1 - \frac{\lambda_n}{n}\right)^{n-k} \\
&= 1 \cdot \left(1 - \frac{1}{n}\right) \cdot \left(1 - \frac{2}{n}\right) \cdots \left(1 - \frac{k-1}{n}\right) \cdot \frac{\lambda_n^k}{k!} \cdot \left(1 - \frac{\lambda_n}{n}\right)^{n-k}.
\end{aligned}$$

Граничен преход $n \to \infty$:
Първата част от множителите:

$$\lim_{n \to \infty} \left(1 - \frac{1}{n}\right) \cdot \left(1 - \frac{2}{n}\right) \cdots \left(1 - \frac{k-1}{n}\right) = 1$$

Втората част от множителите:

$$\frac{\lambda_n^k}{k!} \to \frac{\lambda^k}{k!} \text{ (тъй като } \lambda_n \to \lambda)$$

Третата част от множителите:

$$\left(1 - \frac{\lambda_n}{n}\right)^{n-k} = \left(1 - \frac{\lambda_n}{n}\right)^n \cdot \left(1 - \frac{\lambda_n}{n}\right)^{-k}$$

Знаем, че:

$$\lim_{n \to \infty} \left(1 - \frac{\lambda_n}{n}\right)^n = e^{-\lambda} \text{ (тъй като } \lambda_n \to \lambda)$$

$$\lim_{n \to \infty} \left(1 - \frac{\lambda_n}{n}\right)^{-k} = 1$$

Следователно, комбинирайки резултатите, получаваме:

$$\lim_{n\to\infty} \mathbb{P}(X_n = k) = 1 \cdot \frac{\lambda^k}{k!} \cdot e^{-\lambda} \cdot 1 = \frac{\lambda^k e^{-\lambda}}{k!}$$

Това е точно PMF на поасоновото разпределение!

⊕ $p = 0.01$ вероятност даден човек да е заразен с вирус и $n = 1{,}000$.

$$\mathbb{P}(X = 50) = \binom{1{,}000}{50} \times 0.01^{50} \times 0.99^{950} \underset{\lambda=np=10}{\overset{\text{Pois}}{\approx}} \frac{10^{50}}{50!}e^{-10}$$

$$\mathbb{P}(X = 0) \approx \frac{10^0}{0!}e^{-10} = e^{-10}.$$

Практическо правило

Правило на пръста: Може да използваме поасоново приближение за биномно разпределение, когато:

$$n \geq 20 \text{ и } p \leq 0.05$$

Или по-точно:

$$n \geq 100 \text{ и } np \leq 10$$

⊕ Ситуация: Фабрика произвежда чипове с дефектност 0.5%. В партида от 1000 чипа, каква е вероятността да има точно 3 дефектни?

Точен биномен модел:

$$n = 1000, p = 0.005, k = 3$$

$$\mathbb{P}(X = 3) = \binom{1000}{3}(0.005)^3(0.995)^{997}$$

Това трудно се изчислява на ръка.

Поасоново приближение:

$$\lambda = np = 1000 \times 0.005 = 5$$

$$\mathbb{P}(X = 3) \approx \frac{5^3 e^{-5}}{3!} = \frac{125 \times e^{-5}}{6} \approx \frac{125 \times 0.0067379}{6} \approx 0.1404$$

Реалната биномна вероятност е приблизително 0.1404.

Аналитично доказателство на теоремата чрез PGF:

Нека $X_n \in \mathrm{Bin}(n, p_n)$, тогава $g_{X_n} = (1 - p_n + p_n s)^n$. Ако $g_{X_n}(s) \xrightarrow[n\to\infty]{} g_X(s)$, където $g_X(s) = e^{\lambda s} \cdot e^{-\lambda}$, то може да заключим искания резултат, тъй като $e^{\lambda s} \cdot e^{-\lambda}$ е пораждаща функция на $X \in \mathrm{Pois}(\lambda)$.

Може да вземем $p_n = \dfrac{\lambda}{n} + \dfrac{Q_n}{n}$, като $\dfrac{Q_n}{n}$ клони по-бързо от $\dfrac{\lambda}{n}$ към 0. Тогава:

$$\lim_{n\to\infty} g_{X_n}(s) = \lim_{n\to\infty} \left(1 - \frac{\lambda}{n} - \frac{Q_n}{n} + \underbrace{\frac{\lambda}{n}s + \frac{Q_n}{n}s}_{<O\left(\frac{1}{n}\right)}\right)^n = \lim_{n\to\infty} \left(1 - \frac{\lambda}{n}(1-s)\right)^n = e^{\lambda(s-1)}.$$