数据科学基础

Foundations of Data Science

3.3 离散概率分布

陈振宇
南京大学智能软件工程实验室
www.iselab.cn

伯努利分布

伯努利分布（Bernoulli distribution），又名两点分布或0-1分布，是一个离散型概率分布，为纪念瑞士科学家雅各布·伯努利而命名。伯努利试验是只有两种可能结果的单次随机试验。

【伯努利分布】 对于一个随机试验，如果它的样本空间只包含两个元素，即$\Omega=\{0,1\}$, 我们总能在Ω上定义一个服从伯努利分布(又称0-1分布)的随机变量。它的概率分布为:

$$P(X=1)=p, P(X=0)=q, p+q=1$$

对新生婴儿的性别进行登记，检查产品的质量是否合格，几乎所有决策问题都可以用伯努利分布来描述。伯努利分布是最基本的一种分布。

更多详情，请咨询课程助教。

伯努利分布的矩

$$E(X)=0\cdot(1-p)+1\cdot p=p$$

$$E(X^2)=0^2\cdot(1-p)+1^2\cdot p=p$$

$$Var(X)=E(X^2)-E(X)^2=p-p^2=p(1-p)$$

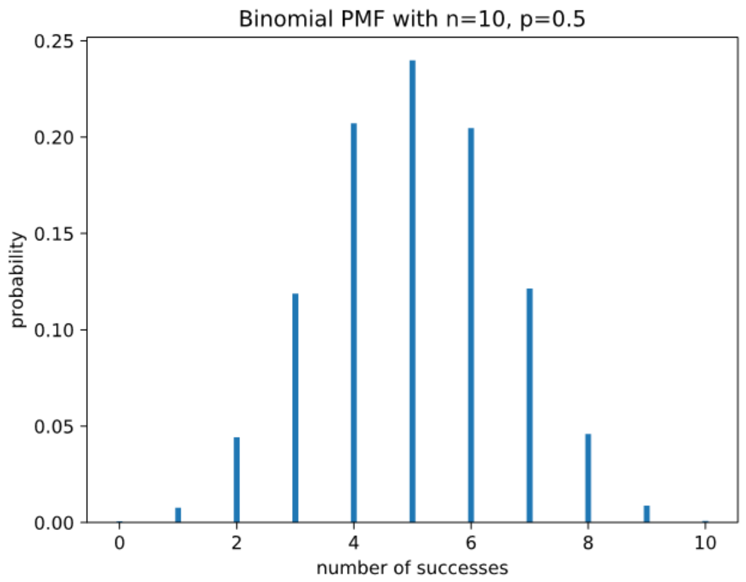
二项分布

二项分布（Binomial Distribution）是n个独立的伯努利试验的离散概率分布。

【二项分布】设事件A在任一次试验中出现的概率为p, 则在n重伯努利试验中事件A发生的次数k的取值为0,,1⋯,n。该随机变量X的概率分布为:

$$P(X=k)=C_{\mathrm{n}}^{\mathrm{k}}\,p^{k}\,(1-p)^{(n-k)}\quad ,k=0,1,\cdots,n$$

则称X服从参数为n, p的二项分布,记为$X\sim\mathbb{B}(n,p)$.



更多详情，请咨询课程助教。

二项分布的矩

设$X=\sum_{i=1}^{n} X_i$，其中X_i服从伯努利分布

$$E(X)=E(\sum_{i=1}^{n} X_i)=\sum_{i=1}^{n} E(Xi)=np$$

$$Var(X)=Var(\sum_{i=1}^{n} X_i)=\sum_{i=1}^{n} Var(Xi)=np(1-p)$$

二项分布示例

【例】某人进行射击, 设每次射击的命中率为0.02, 独立射击400次，试求至少击中两次的概率。

解：将一次射击看成是一次试验. 设击中的次数为X, 则$X\sim\mathbb{B}(400, 0.02)$. X的分布律为

$$P\{X=k\}=C_{400}^{k}(0.02)^{k}(0.98)^{(400-k)}$$

即得所求概率为

$$P\{X\geq2\}=1-P(X=0)-P(X=1)$$

$$=1-(0.98)^{400}-400(0.02)(0.98)^{399}=0.9972$$

一个事件发生的概率不论多小，只要不断重复试验下去，事件迟早会出现的，概率接近1。

二项分布示例

【例】设有80台同类型设备, 各台工作是相互独立的, 发生故障的概率都是0.01. 且一台设备的故障能由一个人处理. 考虑两种配备维修工人的方法, 其一是由4人维护, 每人负责20台; 其二是由3人共同维护80台. 试比较这两种方法在设备发生故障时不能及时维修的概率的大小.

解：按第一种方法. 以X记“第1人维护的20台中同一时刻发生故障的台数”, 以A_i $(i=1,2,3,4)$表示事件“第i人维护的20台中发生故障不能及时维修”, 则知80台中发生故障而不能及时维修的概率为

$P(A_1 \cup A_2 \cup A_3 \cup A_4) \geq P(A\_1) = P\{X \geq 2\}$

而$X \sim \mathbb{B}(20,0.01)$, 故有 $P(A_1 \cup A_2 \cup A_3 \cup A_4) \geq P\{X \geq 2\} = 1 - P\{X=0\} - P\{X=1\} = 0.0169$

按第二种方法. 以Y记80台中同一时刻发生故障的台数. 此时, $Y \sim \mathbb{B}(80,0.01)$, 故80台中发生故障而不能及时维修的概率为 $P\{Y \geq 4\} = 1 \sum_{i=0}^{3} C_{80}^{i} (0.01)^i (0.99)^{(80-i)} = 0.0087$

我们发现, 在后一种情况尽管任务重了(每人平均维护约27台), 但工作效率不仅没有降低, 反而提高了.

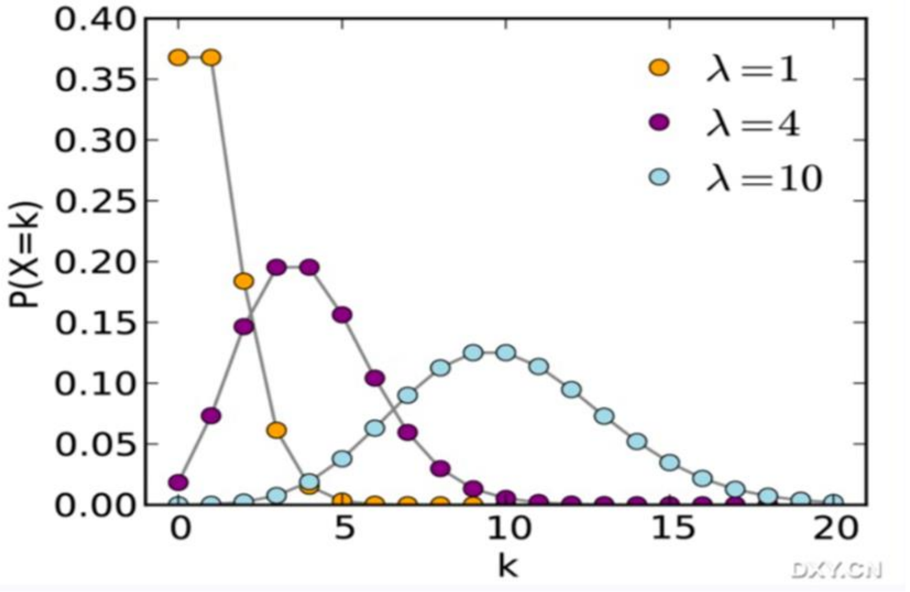
泊松分布

泊松分布（Poisson Distribution）是法国数学家泊松于1837年引入的。泊松分布适合于描述单位时间内随机事件发生的次数的概率分布。如某一服务设施在一定时间内受到的服务请求的次数，系统出现的故障数、自然灾害发生的次数、DNA序列的变异数、放射性原子核的衰变数等等。

【泊松分布】若随机变量X的概率分布为:

$$P(X=k)=\frac{\lambda^k}{k!}e^{-\lambda}, k=0,1,2,\cdots$$

则称X服从参数为λ的泊松分布, 记为$X\sim\pi(\lambda)$.

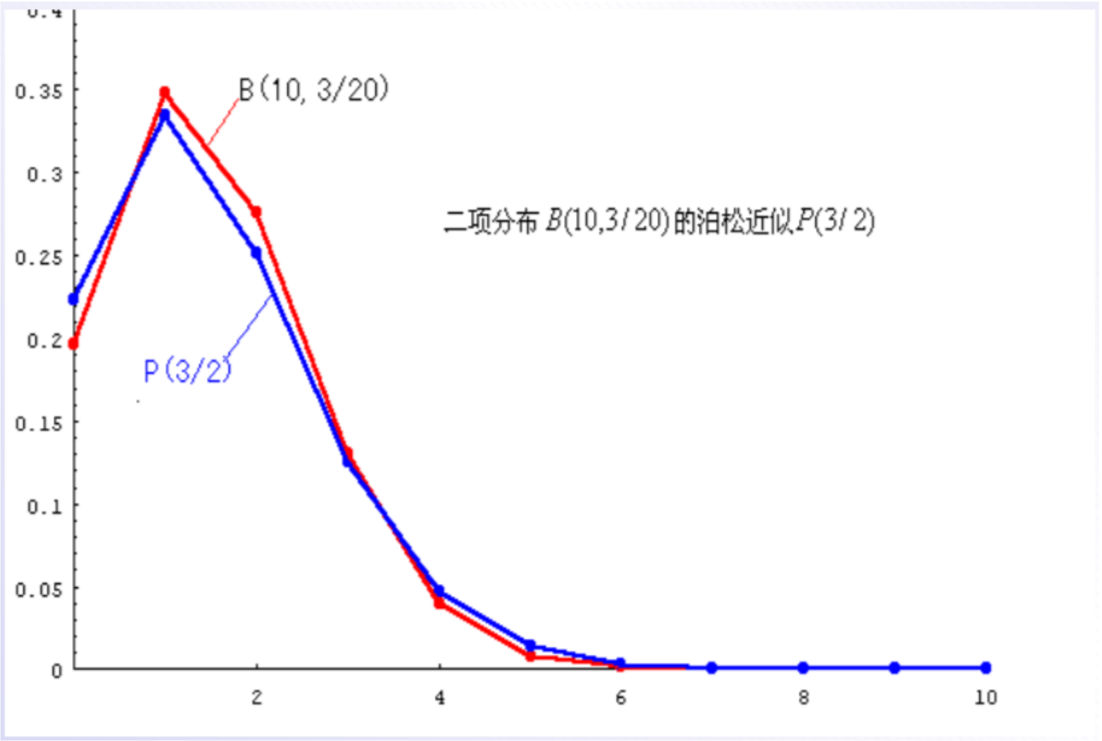


泊松逼近

在很多应用问题中, 我们常常这样的伯努利试验, 其中,相对地说, n大, p小, 而乘积$\lambda = np$大小适中. 在这种情况下, 有一个便于使用的近似公式.

【泊松逼近】 在伯努利试验中, 以p_n代表事件A在试验中出现的概率, 如果$np_n \to \lambda$, 则当$n \to \infty$时,

$$\lim_{n\to\infty} \mathbb{B}\,(n, p_n) = \pi(\lambda)$$



泊松逼近证明

回顾二项分布公式：$P(X=k)=C_n^k p^k(1-p)^{n-k}, p=\frac{\lambda}{n}$

证明：

$$
\begin{aligned}
\lim_{n\to\infty} P(X=k) &= \lim_{n\to\infty} \binom{n}{k} p^k (1-p)^{n-k} \\
&= \lim_{n\to\infty} \frac{n!}{(n-k)!k!} \left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{n-k} \\
&= \lim_{n\to\infty} \underbrace{\left[\frac{n!}{n^k(n-k)!}\right]}_{F} \left(\frac{\lambda^k}{k!}\right) \underbrace{\left(1-\frac{\lambda}{n}\right)^n}_{\to\exp(-\lambda)} \underbrace{\left(1-\frac{\lambda}{n}\right)^{-k}}_{\to 1} \\
&= \lim_{n\to\infty} \underbrace{\left[\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\ldots\left(1-\frac{k-1}{n}\right)\right]}_{\to 1} \left(\frac{\lambda^k}{k!}\right) \underbrace{\left(1-\frac{\lambda}{n}\right)^n}_{\to\exp(-\lambda)} \underbrace{\left(1-\frac{\lambda}{n}\right)^{-k}}_{\to 1} \\
&= \left(\frac{\lambda^k}{k!}\right)\exp(-\lambda)
\end{aligned}
$$

更多详情，请咨询课程助教。

泊松逼近计算

【例】假如生三胞胎的概率为10^{-4}, 求在100000次生育中, 有0, 1, 2次生三胞胎的概率.

解:可看作伯努利试验; $n = 100000, p = 0.0001$, 所求的概率直接计算为

- $\mathbb{B}(0; 100000,0.0001) = 0.000045378$
- $\mathbb{B}(1; 100000,0.0001) = 0.00045382$
- $\mathbb{B}(2; 100000,0.0001) = 0.0022693$

这时也可用泊松逼近, $\lambda = np = 10$, 而

- $\pi(0; 10) = 0.00004540$
- $\pi(1; 10) = 0.0004540$
- $\pi(2; 10) = 0.002270$

泊松逼近计算

【定理】设$X \sim \pi(\lambda)$, 则$E(X) = \lambda$.

证:

$$E(X) = \sum_{k=0}^{\infty} k\lambda^k k!\, e^{-\lambda}$$

$$= \lambda^{-\lambda} \sum_{k=1}^{\infty} \lambda^{k-1}(k-1)! = \lambda e^{-\lambda} e^{\lambda} = \lambda$$

【定理】设$X \sim \pi(\lambda)$, 则$Var(X) = \lambda$.

证:

$$E(X^2) = E[X(X-1) + X] = E[X(X-1)] + E(X)$$

$$= \sum_{k=0}^{\infty} k(k-1)\lambda^k k!\, e^{-\lambda} + \lambda = \lambda^2 e^{-\lambda} \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} + \lambda$$

$$= \lambda^2 e^{-\lambda} e^{\lambda} + \lambda = \lambda^2 + \lambda$$

$$Var(X) = E(X^2) - E(X)^2 = \lambda$$

几何分布

【**例**】从生产线上随机抽产品进行检测,设产品的次品率为p, $0 < p < 1$,若查到次品就停机检修,设停机时已检测到X只产品。X的概率分布律

设A_i为第i个抽到正品事件, A_i相互独立, 则

$$P\{X = n\} = P\{A_1, \cdots, A_{n-1}\overline{A_n}\} = (1-p)^{n-1}p$$

【**几何分布**】 如果随机变量X的概率分布为:

$$P(X = n) = (1-p)^{n-1}p, n = 1,2,\cdots, 0 < p < 1$$

则称X服从几何分布, 记为$X \sim \mathbb{G}(p)$

几何分布的矩

几何分布的数学期望($q=1-p$)：

$$E(X)=\sum_{i=1}^{\infty} ipq^{i-1}$$
$$=1p+2pq+3pq^2+\cdots+kpq^{k-1}+\cdots$$
$$=p\left(1+2q+3q^2+\cdots+kq^{k-1}+\cdots\right)$$

令

$$S=1+2q+3q^2+\cdots+kq^{k-1}+\cdots$$
$$qS=1q+2q^2+3q^3+\cdots+kq^k+\cdots$$

$$S-qS=(1-q)S=1+q+q^2+\cdots+q^k+\cdots=\frac{1}{1-q}$$

所以$S=\frac{1}{(1-q)^2}$,$E(X)=pS=\frac{p}{(1-q)^2}=\frac{1}{p}$

几何分布的矩

几何分布的方差($q=1-p$)：

$$E\left(X^2\right)=\sum_{i=1}^{\infty} i^{2pq^{i-1}}$$

$$=p\left(1+2^{2q}+3^{2q^2}+\cdots+k^{2q^{k-1}}+\cdots\right)$$

令$S=1+2^{2q}+3^{2q^2}+\cdots+k^{2q^{k-1}}+\cdots$, 对$T'=S$, 即$T$关于$q$求一阶导得$S$

$$T=q+2q^2+3q^3+\cdots+kq^k+\cdots=\frac{q}{(1-q)^2}$$

则$S=T'$,

$$S=T'=\frac{(1-q)^2+2(1-q)q}{(1-q)^4}=\frac{1-q^2}{(1-q)^4}=\frac{1+q}{(1-q)^3}$$

所以$E\left(X^2\right)=pS=\frac{2-p}{p^2}$

$$D(X)=E\left(X^2\right)-(E\mathrm{X})^2=\frac{2-p}{p^2}-\frac{1}{p^2}=\frac{1-p}{p^2}$$

几何分布的无记忆性

【无记忆】对于一个非负随机变量X, 如果对于任意$s,t \geq 0$有

$$P\{X > s + t|X > t\} = P\{X > s\}$$

则我们称X是无记忆的。

$$P\{X \leq n\} = 1 - q^n, P\{X > n\} = p^n$$

$$P\{X > n\} = \sum_{i=n+1}^{\infty} pq^{i-1} = pq^n \sum_{i=0}^{\infty} q^i = \frac{pq^n}{1-q} = q^n$$

因此无记忆性是几何分布所具有的一个有趣的性质。

超几何分布

一批产品共N件，含M件是次品，随机地从这N件产品中抽取n件产品，求恰有k件次品的概率。

【超几何分布】 如果随机变量X的概率分布为:

$$P(X=k)=\frac{C_M^k C_{N-M}^{n-k}}{C_N^n}, k=1,2,\cdots,n$$

其中N, M, n均为正整数, 且$M \leq N, n \leq N$,则称X服从参数为N, M, n的超几何分布, 记为$X \sim \mathbb{H}(N, M, n)$.

超几何分布近似

我们把二项分布与超几何分布作一比较。N件产品，有M件次品和$N-M$件正品。如果每抽一件产品放回后，再抽下一件产品，如此有放回地随机地抽取n件，这是n重伯努利试验，那么所抽的n件产品的次品数$X \sim \mathbb{B}\,(n, \frac{M}{N})$。当$N >> n$时，我们可以采用二项分布近似超几何分布。

超几何分布、二项分布和泊松分布都是重要的离散型随机变量的概率分布. 有时, 他们的概率计算会十分繁冗. 当试验次数n很大时,可以推导出这三个分布间有一种近似关系式

$$\frac{C_M^k C_{N-M}^{n-k}}{C_N^n} \approx \mathrm{C}_n^k p^k (1-p)^{n-k} \approx \frac{\lambda^k}{k!} e^{-\lambda}$$

超几何分布的矩

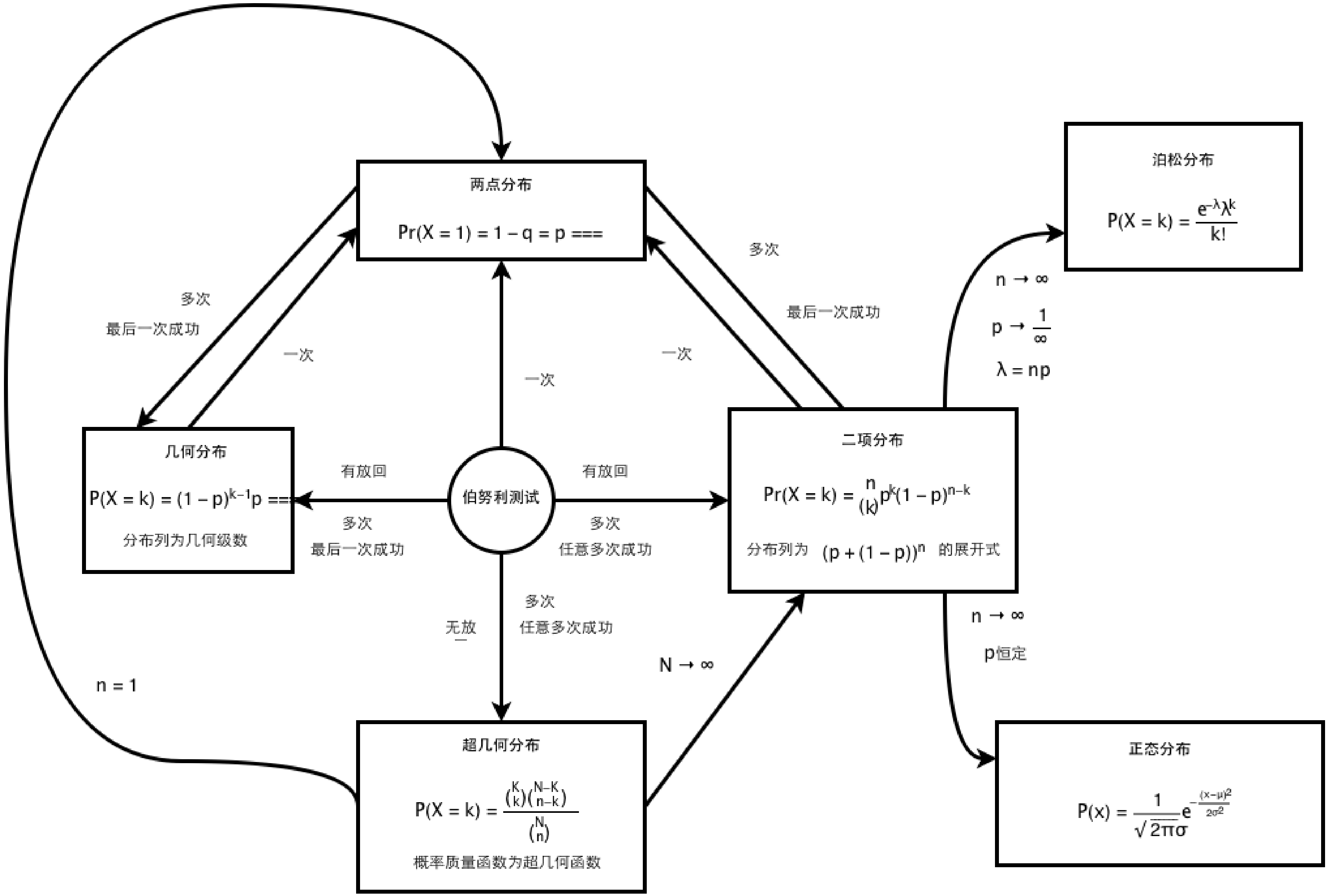
【超几何分布的数学期望】

$$n\frac{M}{N}$$

【超几何分布的方差】

$$n\frac{M}{N}\frac{N-M}{N}\frac{N-n}{N-1}$$

总结



更多详情，请咨询课程助教。

更多详情，请咨询课程助教。