2. 概率论

2.1 排列和组合

1. 摸球问题

（1）一共有 n 个小球，从中选出 m 个进行排列（有顺序），一共有多少种组合?

解：第一次从 n 个小球中抽出一个，第二次从 $n-1$ 个小球中抽出一个，如此类推，一共有：

$$A_n^m = n \times (n-1) \times (n-2) \times \ldots \times (n-m+1) = \frac{n!}{(n-m)!}$$

（2）一共有 n 个小球，从中选出 m 个放到一个盒子里（无顺序），一共有多少种组合?

解：第一次从 n 个小球中抽出一个，第二次从 $n-1$ 个小球中抽出一个，如此类推，一共有：

$$n \times (n-1) \times (n-2) \times \ldots \times (n-m+1) = A_n^m$$

因为盒子中的小球没有顺序，所以还要除以组合情况，m 个小球的组合情况有$m!$，因此：

$$C_n^m = \frac{n \times (n-1) \times (n-2) \times \ldots \times (n-m+1)}{m \times (m-1) \times (m-2) \times \ldots \times 1} = \frac{n!}{(n-m)!m!} = \frac{A_n^m}{A_m^m}$$

2.2 概率知识

$$P(x) \in [0,1]$$

概率为0的事件不一定不会发生，概率为1的事情不一定会发生，比如：投针、连续型随机变量。

举例：一张白纸上有一个黑点，抛一枚硬币落在纸上，硬币落在黑点上的概率为0，但并不表示不可能发生，没有落在黑点上的概率为1，其也不是必然事件。

2.2.1 累积分布函数

$$F(x) = P\left(x \leq x_0\right)$$

$F(x)$ 的值域在$[0,1]$，且一定为单调递增函数。

实际上反过来，我们可以把值域在$[0,1]$的单调递增函数$F(x)$看作是某个事件的 X 累积分布函数。

对应离散型则是累积概率，$\sum_{k=1}^{m} P(x=k)$ 。

2.2.2 概率密度函数

若累积分布函数可导，则定义 $f(x) = F'(x)$ 为概率密度函数。

其对应离散型就是 $P(x=k)$

2.3 概率公式

条件概率公式：$P(A|B) = \frac{P(AB)}{P(B)}$

全概率公式：$P(A) = \sum_{i=1}^{m} P\left(A|B_i\right) P\left(B_i\right)$

贝叶斯公式：

$$P\left(B_i|A\right) = \frac{P\left(B_i A\right)}{P(A)} = \frac{P\left(A|B_i\right) P\left(B_i\right)}{\sum_{j=1}^{m} P\left(A|B_j\right) P\left(B_j\right)}$$

贝叶斯公式的进一步理解：

$$P(\theta|x) = \frac{P(x|\theta)P(\theta)}{P(x)}$$

其中 θ 是系统参数，x 是从系统得到的样本（现象）

$P(\theta)$:先验概率，即没有得到观察时，系统出现的 θ 概率；

$P(\theta|x)$:后验概率，即获得观察 x 后，倒推系统出现 θ 的概率；

$P(x|\theta)$:似然函数，给定系统参数 θ 下的事件概率分布；

举例：8支步枪中有5支已校准过，3支未校准。一名射手用校准过的枪射击，中靶概率为0.8；用未校准的枪射击，中靶概率为0.3；现从8支枪中随机取一支射击，结果中靶。求该枪是已校准过的概率。

解：假设 θ 是校准过的步枪($G=1$)，x 是射击中靶 ($A=1$)，

则有如下规定：

$P(\theta)$:先验概率，从8支枪中选一支，枪是校准过的概率；

$P(\theta|x)$:后验概率，从8支枪中选一支，结果中靶，枪是校准过的概率；

$P(x|\theta)$:似然函数，一支校准过的枪，该枪中靶的概率；

$$\mathrm{P(G=1)}=\frac{5}{8}\quad \mathrm{P(G=0)}=\frac{3}{8}$$

$$\mathrm{P(A=1|G=1)}=0.8\quad \mathrm{P(A=0|G=1)}=0.2$$

$$\mathrm{P(A=1|G=0)}=0.3\quad \mathrm{P(A=0|G=0)}=0.7$$

$$\mathrm{P(G=1|A=1)}=?$$

$$\mathrm{P(G=1|A=1)}=\frac{P(A=1|G=1)P(G=1)}{\sum_{i\in G}P(A=1|G=i)P(G=i)}=\frac{0.8\times\frac{5}{8}}{0.8\times\frac{5}{8}+0.3\times\frac{3}{8}}=0.8163$$

总的来说，贝叶斯公式最大作用是解决逆概问题。

2.4 事件的独立性、期望和方差

2.4.1 独立性定义

给定事件 A 和 B ，事件 A 和 B 相互独立的充要条件为：

$$P(AB)=P(A)P(B)$$

对于随机变量的分布函数，则有：

$$F\left(x_1,x_2\right)=F_1\left(x_1\right)\cdot F_2\left(x_2\right)$$

2.4.2 期望

1. 期望的计算

离散型：$E(X)=\sum_{i=1}^{m}x_ip_i$

连续型：$E(X)=\int_{-\infty}^{\infty}xf(x)dx$

2. 期望的性质

（1） 一般情况：

$E(kX)=kE(X)$ ，k为常数

$E(X+Y)=E(X)+E(Y)$

（2）在独立的前提下：

若 X 和 Y 相互独立，则有 $E(XY)=E(X)E(Y)$

注：上式并不是相互独立的充要条件。在独立的情况下，必有上式成立，而上式成立只能说明 X 和 Y 不相关。

2.4.3 方差

1.方差的计算

$$\mathrm{Var}(X)=E\left\{[X-E(X)]^2\right\}=E\left(X^2\right)-E^2(X)\geq 0$$

上式仅当随机变量 X 为定值时，取得等号。

2.方差的性质

（1） 一般情况：

$$\operatorname{Var}(c)=0\ （c为常数）$$
$$\operatorname{Var}(X+c)=\operatorname{Var}(X)$$
$$\operatorname{Var}(kX)=k^2\operatorname{Var}(X)$$

（2）在独立的前提下：

$$\operatorname{Var}(X+Y)=\operatorname{Var}(X)+\operatorname{Var}(Y)$$

2.4.4 协方差

1. 协方差的计算

$$\operatorname{Cov}(X,Y)=E\{[X-E(X)][Y-E(Y)]\}=E(XY)-E(X)E(Y)$$

2. 协方差的性质

（1） 一般情况：

$$\operatorname{Cov}(X,Y)=\operatorname{Cov}(Y,X)$$
$$\operatorname{Cov}(aX+b,cY+d)=\mathrm{ac}\operatorname{Cov}(X,Y)$$
$$\operatorname{Cov}(X_1+X_2,Y)=\operatorname{Cov}(X_1,Y)+\operatorname{Cov}(X_2,Y)$$
$$\operatorname{Cov}(X,Y)=E(XY)-E(X)E(Y)$$

（2）在独立的前提下：

$$E(XY)=E(X)E(Y)$$
$$\operatorname{Cov}(X,Y)=E(XY)-E(X)E(Y)=0$$

实际上，上式依然不是充要条件。独立情况下有上式成立，上式成立只能说明 X 和 Y 不相关。

3. 协方差的上界

$$|\operatorname{Cov}(X,Y)|\leq\sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}$$

协方差上界定理的证明：

取任意实数 t，构造随机变量 Z，

$$Z=(\mathrm{X}-\mathrm{E}(X))\cdot\mathrm{t}+(\mathrm{Y}-\mathrm{E}(Y))$$

从而：

$$\begin{cases}E\left(Z^2\right)=\sigma_1^2t^2+2\operatorname{Cov}(X,Y)\cdot t+\sigma_2^2\\ E\left(Z^2\right)\geq 0\end{cases}$$

$$\Rightarrow\sigma_1^2t^2+2\operatorname{Cov}(X,Y)\cdot t+\sigma_2^2\geq 0$$
$$\Rightarrow\Delta=4\operatorname{Cov}^2(X,Y)-4\sigma_1^2\sigma_2^2\leq 0$$
$$\Rightarrow|\operatorname{Cov}(\mathrm{X},\mathrm{Y})|\leq\sigma_1\sigma_2$$

4. 相关系数

$$\rho_{xy}=\frac{\operatorname{Cov}(X,Y)}{\sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}}\leq 1$$

当且仅当 X 和 Y 是完全线性相关时，等号成立。

5. 协方差及相关系数的通俗理解：

假设有个时刻，每个时刻都有变量以及变量。从公式可以看到，对于变量以及变量，协方差研究的是每个时刻"变量与其均值(期望)之差"以及"变量与其均值（期望）之差"的乘积，再对这个乘积相加并求平均值（实际也是一个期望）。这个时候反映了什么？反映了变量以及变量是否是做同向移动。如果两个变量在某时刻的变化是同向的，那么乘积则为正值。相反，如果变动是反向的，乘积则为负值。所以协方差其实就是反映了两个变量同向变动的程度；

一个小提示：上面用以及作为对变动方向的研究，其实还是有一点不严谨，因为这个只是反映了相比于均值的波动。

先让我们回看一下相关系数。

$$\rho_{xy} = \frac{\operatorname{Cov}(X,Y)}{\sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}}$$

实际上就是用协方差除以各自的标准差，实际上大家可以把它看做是标准化后的协方差，范围在-1到1之间。

需要注意的是，协方差和相关系数非常受到异常值的影响：

有数据x=[1,-1,1,-1,1,-1,1,-1,100]，y=[-1,1,-1,1,-1,1,-1,1,100]

前8个数据计算相关系数，得到结果为-1.但是加上第9个数据，相关系数则达到了99.8%。

另外，有数据x=[-2,-6,-6,-2,2,6,6,2]以及y=[-6,-2,-2,-6,6,2,2,6]

前面4个数据相关系数为-1，后面4个数据相关系数同样为-1，但是8个数据的相关系数则为0.6

6. 协方差矩阵

对于 n 个随机向量 $(X_1, X_2, \dots X_m)$，任意两个元素和 X_i 和 X_j 以得到一个协方差，从而形成协方差矩阵，其中$c_{ij} = \operatorname{Cov}(X_i, X_j)$

$$C = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1m} \\ c_{21} & c_{22} & \cdots & c_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ c_{m1} & c_{m2} & \cdots & c_{mm} \end{bmatrix}$$

2.5 常见的分布

2.5.1 0-1分布（离散）

1.定义

两点分布(two-point distribution)即"伯努利分布"，也称0-1分布。在一次试验中，事件 A 出现的概率为 P，事件 A 不出现的概率为 $q = 1 - p$，若以 X 记一次试验中 A 出现的次数，则 X 仅取0、1两个值。 X 的概率分布为$P(X = k) = p^k q^{1-k},\ k = 0, 1$，称 X 服从伯努利分布。

X	1	0
P	p	$1 - p$

2.性质

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

$D(X)=E\left(X^2\right)-[E(X)]^2=1^2\times p+0^2\times(1-p)-p^2=p(1-p)$

2.5.2 二项分布（离散）

1.定义

设X_i 是相互独立且服从参数为 p 的伯努利分布，则随机变量服从参数为 n 和 p 的二项分布(也称 n 重伯努利分布)：

$$Y=\sum_{i=1}^{n}X_i$$

2.性质

期望和方差的计算1：

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

$$D(X)=\sum_{i=1}^{n}D\left(X_i\right)=np(1-p)$$

期望和方差的计算2：

X 的分布律为：　$\mathrm{P}\{X=k\}=\binom{n}{k}p^k(1-p)^{n-k},(k=0,1,2,\ldots,n)$

则有

$$\begin{aligned}\mathrm{E(X)}=\sum_{k=0}^{n}k\cdot P\{X=k\}&=\sum_{k=0}^{n}k\binom{n}{k}p^k(1-p)^{n-k}\\&=\sum_{k=0}^{n}\frac{kn!}{k!(n-k)!}p^k(1-p)^{n-k}\\&=\sum_{k=1}^{n}\frac{np(n-1)!}{(k-1)![(n-1)-(k-1)]!}p^{k-1}(1-p)^{(n-1)-(k-1)}\\&=\mathrm{np}\sum_{k=1}^{n}\frac{(n-1)!}{(k-1)![(n-1)-(k-1)]!}p^{k-1}(1-p)^{(n-1)-(k-1)}\\&=\mathrm{np}[p+(1-p)]^{n-1}=\mathrm{np}\end{aligned}$$

$$\begin{aligned}\mathrm{E}\left(\mathrm{X}^2\right)&=E[X(X-1)+X]=E[X(X-1)]+E(X)\\&=\sum_{k=0}^{n}k(k-1)\binom{k}{n}p^k(1-p)^{n-k}+np\\&=\sum_{k=0}^{n}\frac{k(k-1)n!}{k!(n-k)!}p^k(1-p)^{n-k}+np\\&=\mathrm{n(n-1)}p^2\sum_{k=2}^{n}\frac{(n-2)!}{(n-k)!(k-2)!}p^{k-2}(1-p)^{(n-2)-(k-2)}+np\\&=\mathrm{n(n-1)p}^2[p+(1-p)]^{n-2}+np=\left(n^2-n\right)p^2+np\end{aligned}$$

$$\begin{aligned}\mathrm{D(X)}=&\mathrm{E}\left(\mathrm{X}^2\right)-[\mathrm{E(X)}]^2=\left(n^2-n\right)p^2+np-(np)^2\\&=\mathrm{np(1-p)}\end{aligned}$$

2.5.3 负二项分布（离散）

对于一系列独立，且服从发生概率为 p 伯努利试验，试验直到第 r 次成功则结束，那试验次数 X 的概率为：

$$P(X=x;r,p)=C_{x-1}^{r-1}\cdot p^r\cdot(1-p)^{x-r},x\in[r,r+1,r+2,\ldots)$$

2.5.4 泊松分布（离散）

1. 定义

泊松分布可以看作一个n很大，而p很小的二项分布，可以理解为是二项分布在连续时间内的的一个极限形式。举个例子，我们说抛硬币，抛了10次，其中硬币朝上有8次的概率是

$$\mathrm{P}\{X=8\}=\binom{10}{8}p^8(1-p)^2,(k=0,1,2,\ldots,n)$$

其中，可以看到我们是把一个试验划分为10次（10个阶段），求事件发生k次的概率。那么如果把这个伯努利试验放在一个连续时间内进行，求在指定时间内事件发生k次的概率，又该如何进行？

不妨先假定指定时间内发生事件的期望为μ,把时间划分为n份，那么在每个试验内（每份时间）事件发生的概率为$p=\frac{\mu}{n}$,那么就有：

$$\begin{aligned}&\lim_{n\to\infty}\binom{n}{k}p^k(1-p)^{n-k}=\lim_{n\to\infty}\binom{n}{k}(\frac{\mu}{n})^k(1-\frac{\mu}{n})^{n-k}\\&=\lim_{n\to\infty}\frac{n(n-1)(n-2)\cdots(n-k+1)}{k!}\frac{\mu^k}{n^k}(1-\frac{\mu}{n})^{n-k}\\&=\lim_{n\to\infty}\frac{\mu^k}{k!}\frac{n}{n}\frac{n-1}{n}\cdots\frac{n-k+1}{n}(1-\frac{\mu}{n})^{-k}(1-\frac{\mu}{n})^n\end{aligned}$$

其中：

$$\begin{aligned}\lim_{n\to\infty}\frac{n}{n}\frac{n-1}{n}\cdots\frac{n-k+1}{n}&=1\\\lim_{n\to\infty}(1-\frac{\mu}{n})^{-k}&=1\\\lim_{n\to\infty}(1-\frac{\mu}{n})^n=\lim_{n\to\infty}(1+\frac{1}{\frac{n}{-\mu}})^{\frac{n}{-\mu}(-\mu)}&=e^{-\mu}\end{aligned}$$

得到有

$$\lim_{n\to\infty}\binom{n}{k}p^k(1-p)^{n-k}=\frac{\mu^k}{k!}e^{-\mu}$$

一般地，在泊松分布中，我们把μ写作λ,有：

$$\lim_{n\to\infty}\binom{n}{k}p^k(1-p)^{n-k}=\frac{\lambda^k}{k!}e^{-\lambda}$$

好了，以上我们得到了泊松分布的公式，我们说，假如一个随机变量X，智能取得非负整数，且它对一个的概率分布为$\frac{\lambda^k}{k!}e^{-\lambda}$,那么我们可以把其称之为一个泊松过程。实际上，泊松过程还有定义如下：

（1）对时间段无限划分为无限个小时间段，在每一份无限接近于零的时间段内，时间发生一次的概率与这个极小时间段的长度成正比；

（2）在划分的无限小的时间段内，事件发生两次及两次以上的概率恒等于零；

（3）在不同的时间段内，事件发生与否相互独立；

常见的泊松分布有：

- 某一服务设施在一定时间内的到达人数

- 电话交换机接到呼叫的次数
- 汽车站台的侯客人数
- 机器出现的故障数
- 自然灾害发生的次数
- 一块产品的缺陷数
- 显微镜下单位分区内的细菌分布数
- 某放射性物质在单位时间内发射出的粒子数

例如我们知道某日料店平均每小时能够卖出5碗鳗鱼饭，那么某个小时，恰好卖出8碗鳗鱼饭的概率，则可以把$\lambda = 5, k = 8$带入公式得到6.53%。

$$\frac{5^8}{8!}e^{-5} = 0.0653$$

当然，我们也可以算算，这个日料店，某个小时一碗鳗鱼饭也卖不出去的概率是0.674%：

$$\frac{5^0}{0!}e^{-5} = 0.00674$$

特别地，我们提到泊松分布研究的是单位时间内发生次数的研究，那么如果要研究给定某段时间内的发生概率则有：

$$P(N(t) = k) = \frac{(\lambda t)^k}{k!}e^{-\lambda t}$$

那么2个小时，这个日料店恰好卖出8碗鳗鱼饭的概率有11.3%：

$$P(N(2) = 8) = \frac{(5 \times 2)^8}{8!}e^{-5 \times 2} = 0.113$$

(关于泊松分布，另一个关键是指数分布)

2.另一种推导（仅供参考）

泰勒展开式有：

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots + \frac{x^n}{n!} + o\left(x^n\right)$$

$$1 = 1 \cdot e^{-x} + x \cdot e^{-x} + \frac{x^2}{2!} \cdot e^{-x} + \frac{x^3}{3!} \cdot e^{-x} + \ldots + \frac{x^n}{n!} \cdot e^{-x} + o\left(x^n\right) \cdot e^{-x}$$

我们现在不考察 x ，将其作为参数：

$$\frac{x^k}{k!} \cdot e^{-x} \to \frac{\lambda^k}{k!} \cdot e^{-\lambda}$$

设$X \sim \pi(\lambda)$，其分布律为

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

3.性质

$$E(X) = \sum_{k=0}^{\infty} k \cdot \frac{\lambda^k}{k!}e^{-\lambda} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^{k-1}}{(k-1)!}\lambda = e^{-\lambda} \cdot e^{\lambda} \cdot \lambda = \lambda$$

（上式利用了泰勒展开）

$$
\begin{aligned}
\mathrm{E}\left(\mathrm{X}^2\right) &= \mathrm{E}[\mathrm{X}(\mathrm{X}-1)+\mathrm{X}] \\
&= \mathrm{E}[\mathrm{X}(\mathrm{X}-1)]+\mathrm{E}(\mathrm{X}) \\
&= \sum_{k=0}^{+\infty} k(k-1)\cdot\frac{\lambda^k}{k!}\mathrm{e}^{-\lambda}+\lambda \\
&= \lambda^2 e^{-\lambda}\sum_{k=2}^{+\infty}\cdot\frac{\lambda^{k-2}}{(k-2)!}+\lambda=\lambda^2 e^{-\lambda}e^{\lambda}+\lambda=\lambda^2+\lambda
\end{aligned}
$$

所以

$$
D(X)=E\left(X^2\right)-[E(X)]^2=\lambda^2+\lambda-\lambda^2=\lambda
$$

2.5.5 均匀分布（连续）

1.定义

设 $X\sim U(a,b)$，其概率密度函数为：

$$
f(x)=\begin{cases}\frac{1}{b-a}, x\in[a,b] \\ 0, \quad \text{其他}\end{cases}
$$

2.性质

$$
\begin{aligned}
E(X) &= \int_{-\infty}^{\infty} x f(x)dx=\int_a^b x\cdot\frac{1}{b-a}dx=\frac{1}{2}(a+b) \\
D(X) &= E\left(X^2\right)-E^2(X)=\int_a^b x^2\cdot\frac{1}{b-a}dx-\left(\frac{a+b}{2}\right)^2 \\
&= \frac{1}{3}\left(a^2+b^2+2ab\right)-\frac{1}{4}(a+b)^2=\frac{(b-a)^2}{12}
\end{aligned}
$$

2.5.6 指数分布（连续）

1.定义

前面我们谈到，泊松分布研究的是单位时间内时间发生次数的。那么指数分布则是研究时间发生的时间间隔概率，如：

- 婴儿出生时间的间隔
- 鳗鱼饭卖出的实践间隔
- 机器出现故障的时间间隔
- 客服中心呼入电话的时间间隔

举个例子：假定某日料店平均每个单位时间能卖出5碗鳗鱼饭。如果两次卖出下鳗鱼饭要间隔t个时间，那么也就意味着在t个时间内，没有卖出一碗鳗鱼饭，有：

$$
P(X>t)=P(N(t)=0)=\frac{(\lambda t)^0}{0!}e^{-\lambda t}=e^{-\lambda t}
$$

因此两次卖出鳗鱼饭的间隔时间小于t，就有X的累积分布函数：

$$
P(X\leq t)=1-P(X>t)=1-e^{-\lambda t}
$$

求导可得：

$$p(x)=\begin{cases}\lambda e^{-\lambda x}, x>0\\ 0, x\leq 0\end{cases}$$

2.性质

$$\begin{aligned}\mathrm{E(X)}&=\int_{-\infty}^{+\infty}xf(x)dx=\int_{0}^{+\infty}x\cdot\lambda\mathrm{e}^{-\lambda x}\mathrm{dx}\\&=-\mathrm{xe}^{-\lambda x}\Big|_0^{+\infty}+\int_0^{+\infty}e^{-\lambda x}dx=\frac{1}{\lambda}\\\mathrm{D(X)}&=\mathrm{E}\left(\mathrm{X}^2\right)-[\mathrm{E(X)}]^2=\int_0^{+\infty}x^2\cdot\lambda\mathrm{e}^{-\lambda x}dx-\theta^2\\&=2\frac{1}{\lambda^2}-\frac{1}{\lambda^2}=\frac{1}{\lambda^2}\end{aligned}$$

3.特性

- 其中 $\lambda>0$是分布的一个参数，常被称为率参数 (rate parameter)。这个定义和泊松分布一致，即每单位时间内发生某事件的次数。
- 指数分布可以用来表示独立随机事件发生的时间间隔，比如旅客进机场的时间间隔、软件更新的时间间隔等等。许多电子产品的寿命分布一般服从指数分布。有的系统的寿命分布也可以用指数分布来近似。它在可靠性研究中是最常用的一种分布形式。
- 可以看到指数分布的期望是$\frac{1}{\lambda}$,如果λ越大，说明间隔时间越短。
- 指数分布的区间是 $[0,\infty)$。如果一个随机变量 X 呈指数分布，则可以写作：X~Exponential（λ）。
- 指数函数的一个重要特征是无记忆性（遗失记忆性，Memoryless Property）。

  如果一个随机变量呈指数分布，当 $s,t\geq 0$ 时有：

$$P(x>s+t|x>s)=P(x>t)$$

即，如果 x 是某电器元件的寿命，已知元件使用了 s 小时，则共使用至少 $s+t$ 小时的条件概率，与从未使用开始至少 t 小时的概率相等。

4.另一种表达

设随机变量 X 服从指数分布，其概率密度为：

$$f(x)=\begin{cases}\frac{1}{\theta}e^{-\frac{x}{\theta}}, x>0\\ 0\quad\ , x\leq 0\end{cases}$$

$$\begin{aligned}\mathrm{E(X)}&=\int_{-\infty}^{+\infty}xf(x)dx=\int_{0}^{+\infty}x\cdot\frac{1}{\theta}\mathrm{e}^{-x/\theta}\mathrm{dx}\\&=-\mathrm{xe}^{-\frac{x}{\theta}}\Big|_0^{+\infty}+\int_0^{+\infty}e^{-\frac{x}{\theta}}dx=\theta\\\mathrm{D(X)}&=\mathrm{E}\left(\mathrm{X}^2\right)-[\mathrm{E(X)}]^2=\int_0^{+\infty}x^2\cdot\frac{1}{\theta}\mathrm{e}^{-\frac{x}{\theta}}dx-\theta^2\\&=2\theta^2-\theta^2=\theta^2\end{aligned}$$

或者也可以写为如下形式，那对应的期望和方差则是$\frac{1}{\lambda}$以及$\frac{1}{\lambda^2}$

$$f(x)=\begin{cases}\lambda e^{-\lambda x} & x>0\\ 0 & x\leq 0\end{cases}$$

2.5.7 正态分布

1.定义

设 $X \sim N\left(\mu, \sigma^2\right)$，其概率密度函数为：

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \sigma>0, x \in(-\infty, \infty)$$

2.性质

$$\begin{aligned}\mathrm{E}(\mathrm{X}) &= \int_{-\infty}^{+\infty} x f(x) dx \\ &= \int_{-\infty}^{+\infty} x \cdot \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx\end{aligned}$$

令 $\frac{x-\mu}{\sigma}=t \Rightarrow x=\mu+\sigma t$

$$\begin{aligned}\mathrm{E}(\mathrm{X}) &= \int_{-\infty}^{+\infty} x \cdot \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \mathrm{dx} \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} (\mu+\sigma t) e^{-\frac{t^2}{2}} dt \\ &= \mu \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{t^2}{2}} dt + \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} t e^{-\frac{t^2}{2}} dt \\ &= \mu \cdot \frac{1}{\sqrt{2}\pi} \cdot 2 \cdot \sqrt{2} \cdot \frac{\pi}{2} + 0 \\ &= \mu\end{aligned}$$

注意此处利用可超越积分：

$$\int_0^{+\infty} e^{-x^2} \mathrm{d}x = \frac{\sqrt{\pi}}{2}$$
$$\int_0^{+\infty} e^{\frac{-x^2}{2}} \mathrm{d}x = \frac{\sqrt{2\pi}}{2}$$

$$\begin{aligned}\mathrm{D}(\mathrm{X}) &= \int_{-\infty}^{+\infty} (x-\mu)^2 f(x) dx \\ &= \int_{-\infty}^{+\infty} (x-\mu)^2 \cdot \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx\end{aligned}$$

令 $\frac{x-\mu}{\sigma}=t$，得

$$\begin{aligned}\mathrm{D}(\mathrm{X}) &= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} t^2 e^{-\frac{t^2}{2}} dt \\ &= \frac{\sigma^2}{\sqrt{2\pi}} \left(-t e^{-\frac{t^2}{2}} \Big|_{-\infty}^{+\infty} + \int_{-\infty}^{+\infty} e^{-\frac{t^2}{2}} dt \right) \\ &= 0 + \frac{\sigma^2}{\sqrt{2\pi}} \sqrt{2\pi} = \sigma^2\end{aligned}$$

2.5.8 Beta分布

1.$Beta$分布的定义：

$Beta$ 分布的概率密度：

$$f(\mathrm{x}) = \begin{cases} \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}, & x \in [0.1] \\ 0 & ,\text{其他} \end{cases}$$

其中系数 B 为：

$$\mathrm{B}(\alpha,\beta)=\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\mathrm{dx}=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

$Gamma$ 函数可以看成阶乘的实数域推广：

$$\Gamma\mathrm{x}=\int_0^\infty t^{x-1}\mathrm{e}^{-t}\mathrm{dt}\Rightarrow\Gamma(\mathrm{n})=(\mathrm{n}-1)!\Rightarrow\mathrm{B}(\alpha,\beta)=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

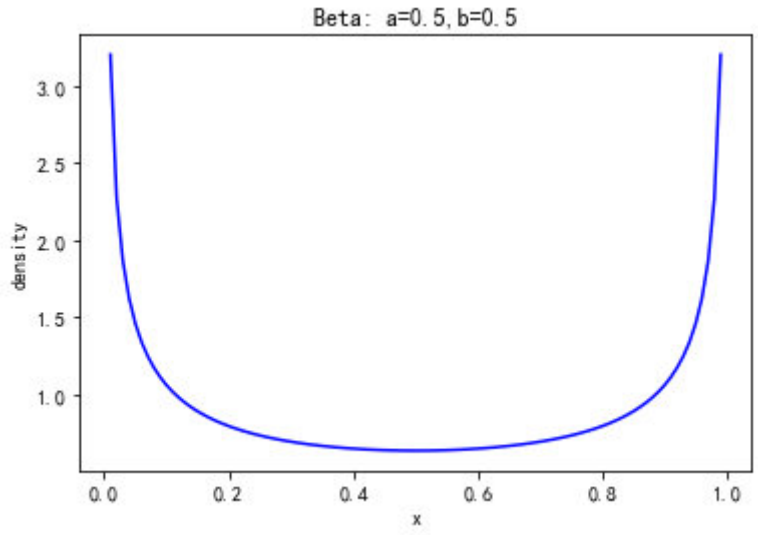
$Beta$ 分布的期望:

$$f(\mathrm{x})=\frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1},x\in[0.1]$$

$$\mathrm{B}(\alpha,\beta)=\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\mathrm{dx}=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

根据定义：

$$\begin{aligned}\mathrm{E(X)}&=\int_0^1 x\cdot\frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}dx\\&=\frac{1}{B(\alpha,\beta)}\int_0^1 x^{(\alpha+1)-1}(1-x)^{\beta-1}dx\\&=\frac{B(\alpha+1,\beta)}{B(\alpha,\beta)}=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\Big/\frac{\Gamma(\alpha+\beta+1)}{\Gamma(\alpha+1)\Gamma(\beta)}\\&=\frac{\alpha}{\alpha+\beta}\end{aligned}$$



2.5.9 指数族

定义：如果概率密度函数$f(x;\theta)$ 可以写成以下形式，

$$f(x;\theta)=c(\theta)\exp\left\{\sum_{j=1}^{k}c_j(\theta)T_j(x)\right\}h(x)$$

其中， $c(\theta)$，$c_j(\theta)$ 不含 x ，（$c(\theta)$，$c_j(\theta)$ 是不同的项，不要认为后者是前者的分量），$T_j(x)$，$h(x)$ 不含 θ 。分布的支撑$x|f(x;\theta)>0$ 不依赖于参数θ （显然均匀分布不是指数族分布），则称该分布为指数族分布。

指数族分布的标准形式：

可以看到 $c(\theta)$，$c_j(\theta)$ 中均有 θ，可以令 $w_j = c_j(\theta), j = 1, \ldots, k$ 解出 $\theta = \theta(w_1, \ldots, w_k)$，从而 $c(\theta) = c(\theta(w)) = c^*(w)$，即 $c(\theta)$，$c_j(\theta)$ 均用 $(w_1, \ldots, w_k)$ 向量表示。代回密度函数得到指数族的标准形式。

$$f(x;\theta) = c^*(w)\exp\left\{\sum_{j=1}^{k} w_j T_j(x)\right\}h(x)$$

例子：

1.二项分布

$$P_\theta(x) = \binom{n}{x}\theta^x(1-\theta)^{n-x} = \binom{n}{x}(1-\theta)^n \exp\left\{x\ln\frac{\theta}{1-\theta}\right\}$$

其中，$c(\theta)$ 为 $(1-\theta)^n$，$h(x)$ 为 $\binom{n}{x}$，$c_j(\theta)$ 为 $ln\frac{\theta}{1-\theta}$，$T_j(\theta)$ 为 x。

不管原形式是怎样，只要能化成指数族分布的形式即为指数族分布。

2.正态分布

$$P_\theta(x) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \left[\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{\mu^2}{2\sigma^2}}\right]e^{-\frac{x^2}{2\sigma^2}+\frac{\mu}{\sigma^2}x}$$
$$= c(\mu,\sigma)\exp\left\{c_1(\mu,\sigma)x + c_2(\mu,\sigma)x^2\right\}$$

$$T_1(x) = x, T_2(x) = x^2, h(x) = 1$$

再把它化成标准指数族分布看看，

令，$w_1 = c_1 = \frac{\mu}{\sigma^2}, w_2 = c_2 = -\frac{1}{2\sigma^2}$，得 $\mu = -\frac{w_1}{2w_2}, \sigma = \sqrt{-\frac{1}{2w_2}}$。

换掉 $c(\mu,\sigma), c_1(\mu,\sigma), c_2(\mu,\sigma)$，得

$$P_{(w_1,w_2)} = \sqrt{\frac{-w_2}{\pi}}e^{\frac{w_1^2}{4w_2}}e^{w_1x+w_2x^2}$$

正态分布原始形式化成指数族形式，再化为指数族标准形式（用 $w = (w_1, w_2)$ 表示），参数均为两个，本质都是正态分布，只是形式不同。

定理1：

$T(X)$ 是 θ 的充分完备统计量，$\hat{g}(X)$ 是 $g(\theta)$ 的无偏估计且方差有限，则 $\hat{g}|T$ 是 $g(\theta)$ 的 UMVUE。($\hat{g}|T$ 指 $\hat{g}(X)$ 是由 $T(X)$ 构成。)

定理2：

指数族分布抽取 n 个 id 的样本 $X = \{X_1, \ldots, X_n\}$，联合概率分布可以写为，

$$f(X;\theta) = c(w)\exp\left\{\sum_{j=1}^{k} w_j T_j(X)\right\}h(X)$$

其中，$c(w) = [c^*(w)]^n$，$T_j(X) = \sum_{i=1}^{n} T_j(X_i)$，$h(X) = \prod_i^n h(X_i)$，像求似然把所有样本概率直接相乘即可。则统计量 $(T_1(X), \ldots, T_k(X))$ 为充分完备统计量。

结合定理1和定理2，可以看到指数族分布的联合分布的统计量充分完备统计量，再构造其无偏估计则为 $UMVUE$。这也是指数族分布应用广泛的原因。

利用定理2可以看到正态分布中 $(T_1(X), T_2(X)) = \left(\sum X_i, \sum X_i^2\right)$ 为重复完备统计量，又有

$E\left(\frac{1}{n}T_1\right)=\mu, E\left(\frac{1}{n-1}\left(T_2-\frac{T_1^2}{n}\right)\right)=\sigma^2$ ，再利用定理1有，$\frac{1}{n}T_1=\overline{X}, \frac{1}{n-1}\left(T_2-\frac{T_1^2}{n}\right)=S_n^2$ 为均值方差的 $UMVUE$ 。

性质1：

$$E\left(\sum T_j\left(x_i\right)\right)=-\frac{\partial \ln c(w)}{\partial w_j}$$

$$\operatorname{cov}\left(\sum T_j\left(x_i\right), \sum T_k\left(x_i\right)\right)=-\frac{\partial^2 \ln c(w)}{\partial w_j \partial w_k}$$

注意 $c(w)=[c^*(w)]^n$ 。这个性质可以轻易求出充分完备统计量的期望、协方差、方差。

注：参考https://zhuanlan.zhihu.com/p/66777539

2.6 马尔科夫不等式及切比雪夫不等式

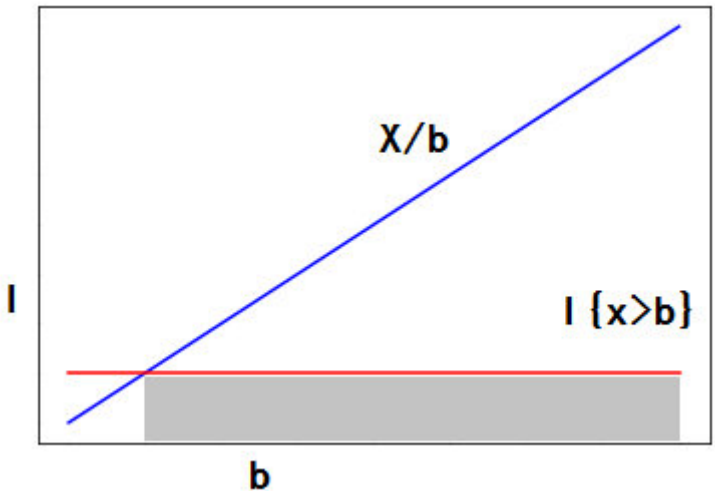
2.6.1 马尔科夫不等式

设一非负随机变量 X，有 $P(X \geq r) \leq \frac{E(X)}{r}$

证明：几何证明方法

在给出直观证明之前，我们需要引入示性函数（indicatorfunction）$I_{(A)}$，示性函数只有在事件 A 成立时才返回1，否则为 0。我们需要使用的一个引理是，当事件是以下形式，如 $A=\{X>a\}$ 时，示性函数的期望可以表示事件发生的概率，即 $\mathrm{E}\left(\mathrm{I}_{(A)}\right)=P(\mathrm{A})$ 。当然， A 的形式不仅限于 $A=\{X>a\}$，$\mathrm{A}=\{X<a\}$ 或 $\mathrm{A}=\{|X|>a\}$ 都是成立的。

引入示性函数后，我们就可以把概率问题转移到更直观的空间上。举例而言，我们知道，对于任意的非负 x 和 b，$\mathrm{I}_{\{\mathrm{x} \geq \mathrm{b}\}} \leq \frac{x}{b}$ 从几何角度而言，即 $\frac{x}{b}$ 在第一象限永远不低于 $\mathrm{I}_{\{\mathrm{x} \geq \mathrm{b}\}}$，如图所示。



因此，将自变量 x 变为随机变量 X ，以上不等式也成立，再对不等式两边取均值，我们可以得到 $Markov$ 不等式，即 $P(X \geq b) \leq \frac{E(X)}{b}$

2.6.3 切比雪夫不等式

设随机变量 X 的期望为 μ，方差为 σ^2，对于任意正数 $\mathcal{E}$，有：

$$P\{|X-\mu| \geq \varepsilon\} \leq \frac{\sigma^2}{\varepsilon^2}$$

上式直接含义就是 X 的波动范围总是集中在均值附近，并且偏移的距离的概率受到方差影响，存在上限。

证明方法1：

把 $|X-\mu|$ 代入马尔科夫不等式，有

$$P(|X-\mu| \geq \varepsilon) \leq \frac{E(|X-\mu|)}{\varepsilon} \Rightarrow P\left((X-\mu)^2 \geq \varepsilon^2\right) \leq \frac{E\left((X-\mu)^2\right)}{\varepsilon^2} = \frac{\sigma^2}{\varepsilon^2}$$

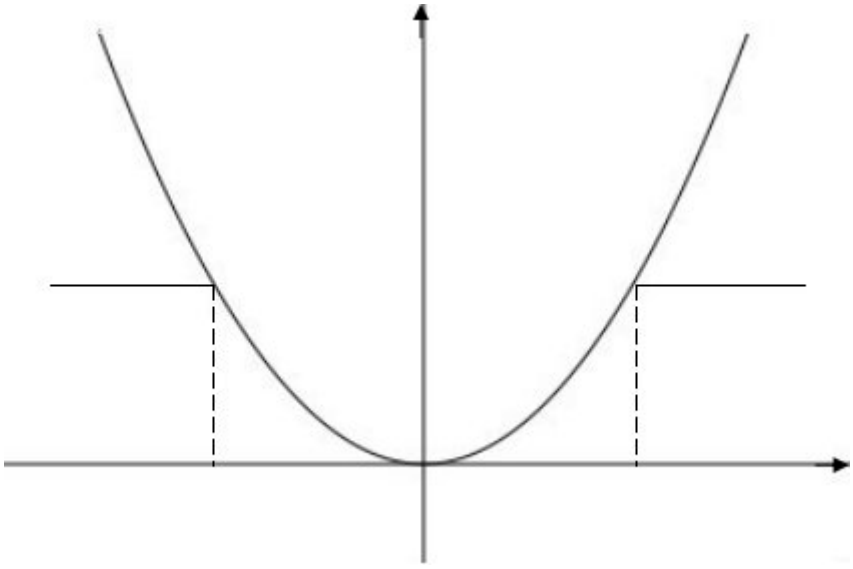
令 $k=\frac{\varepsilon}{\sigma}$，容易得到 k>0：

$$\mathrm{P}(|\mathrm{X}-\mu| \geq \mathrm{k}\sigma) \leq \frac{1}{k^2}$$

证明方法2：图形

对于$Chebyshev$不等式，我们也可以用类似的示性函数来几何直观证明。

对于任意的 x，a 和 b， $I_{(|k-a| \geq b)} \leq \frac{(x-a)^2}{b^2}$ 。右半部分是一个二次函数，而左边是两端取1的示性函数，这个不等式可能难以直接想象出来，不过我们可以画出它的几何形状，从而得到更直观的感觉。



如图所示，二次函数的值在任意点都不会低于示性函数。其中，两坐标轴的交点其实为 $(a,0)$，而两条虚线对应的 x 轴的值分别为 $a-b$ 与 $a+b$，重复上述讨论，将自变量 x 变为随机变量 X，以上不等式也成立，再对不等式两边取均值，同时 a 选择为随机变量 X 的均值 $E(X)$，我们可以得到$Chebyshev$不等式：

$$\mathrm{P}(|\mathrm{X}-\mu| \geq \mathrm{b}) \leq \frac{\mathrm{Var}(X)}{b^2}$$

证明方法3：

$$\mathrm{P}\{|X-\mu| \geq \varepsilon\} = \int_{|X-\mu| \geq \varepsilon} f(x)dx \leq \int_{|X-\mu| \geqslant \varepsilon} \frac{|X-\mu|^2}{\varepsilon^2} f(x)dx$$

$$= \frac{1}{\varepsilon^2} \int_{|X-\mu| \geq \varepsilon} (X-\mu)^2 f(x)dx \leq \frac{1}{\varepsilon^2} \int_{-\infty}^{+\infty} (X-\mu)^2 f(x)d$$

$$= \frac{\sigma^2}{\varepsilon^2}$$

$$\mathrm{P}\{|X-\mu|<\varepsilon\} = 1-\mathrm{P}\{|X-\mu|\geq\varepsilon\} \geq 1-\frac{\sigma^2}{\varepsilon^2}$$

$Chebyshev$的意义及应用：

$Chebyshev$ 的界究竟有多差？我们可以拿正态分布来举个例子。一般来说，正态分布超过两个标准差的概率在约5%左右，而利用 $Chebyshev$ 不等式我们只能说明超过两个标准差的概率一定小于25%，也就是说，利用 $Chebyshev$ 不等式，我们估计随即从正态分布取100个点，平均而言，超过两个标准差的点应该小于25个，而实际上大概只有5个。因此 $Chebyshev$ 的界的确不尽如人意。

但是，它给出了及其偏离的概率上界，这点对证明概率收敛定值非常有效。举例而言，如果我们想证明松散的强大数定律（假设方差存在，实际并不需要这一条件）。假设需要证明如果 X_i 独立同分布于均值为 μ 方差为σ^2 的分布，那么 $\frac{S_n}{n} \to \mu$ 。我们可以用$Chebyshev$不等式，因为 $0 \leq \mathrm{P}\left(\left|\frac{S_n}{n}-\mu \geq \varepsilon\right|\right) \leq \frac{\mathrm{Var}(X)/n}{\varepsilon^2}$ 。当 ε 固定时，$\mathrm{n} \to \infty$，会让 $\mathrm{P}\left(|\frac{S_n}{n}-\mu \geq \varepsilon\right) | \to 0$ 。即证明了方差存在下的强大数定理。（大数定理并不要求方差存在，不过这个证明会更加复杂。）

注：参考https://www.zhihu.com/question/27821324/answer/80814695

2.7 大数定理及中心极限定理

2.7.1 大数定理

N 个具有同期望和方差的独立随机变量的平均值接近原有期望，频率趋向于概率。

设随机变量 $X_1, X_2, \ldots X_n \ldots$ 互相独立，并且具有相同的期望 μ 和方差 σ^2，取前 n 个随机变量的平均

$\mathrm{Y}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$

则对任意正数 ε，有

$$\lim_{n\to\infty} P\{|Y_n-\mu|<\varepsilon\} = 1$$

- 当 n 很大时，随机变量 $X_1, X_2, \ldots X_n$ 的平均值 Y_n 在概率意义下无限接近期望 μ。
  出现偏离是可能的，但这种可能性很小，当 n 无限大时，这种可能性的概率为0。
- 如何证明大数定理?

  提示：根据 Y 的定义，求出它的期望和方差，代入切比雪夫不等式即可。

  一次试验中事件 A 发生的概率为 P；重复 n 次独立试验中，事件 A 发生了 n_A 次，则 p、n、n_A 的关系满足：

对于任意正数 ε，

$$\lim_{n\to\infty} P\left\{\left|\frac{n_A}{n}-p\right|<\varepsilon\right\} = 1$$

上述推论是最早的大数定理的形式，称为**伯努利定理** 。该定理表明事件 A 发生的频率 $\frac{n_A}{n}$ 以概率收敛于事件 A的概率 P，以严格的数学形式表达了频率的稳定性。

上述事实为我们在实际应用中用频率来估计概率提供了一个理论依据。

- 正态分布的参数估计
- 朴素贝叶斯做垃圾邮件分类
- 隐马尔可夫模型有监督参数学习

2.7.2 中心极限定理

中心极限定理是说：样本的平均值约等于总体的平均值。不管总体是什么分布，任意一个总体的样本平均值都会围绕在总体的整体平均值周围，并且呈正态分布，即独立随机变量均值服从正态分布。

设随机变量$X_1, X_2, \ldots X_n \ldots$互相独立，并且具有相同的期望 μ 和方差 σ^2，则随机变量的分布收敛到标准正态分布。

$$\mathrm{Y}_n = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma}$$

容易得到：$\sum_{i=1}^{n} X_i$ 收敛到正态分布 $\mathrm{N}\left(\mathrm{n}\mu, \mathrm{n}\sigma^2\right)$

举例：标准的中心极限定理的问题

有一批样本（字符串），其中 a-z 开头的比例是固定的，但是量很大，需要从中随机抽样。样本量 n，总体中 a 开头的字符串占比1%，需要每次抽到的 a 开头的字符串占比（0.99%，+1.01%），样本量 n 至少是多少?

上述问题也可以重新表述为：大量存在的两点分布 $Bi(1,p)$，其中，Bi 发生的概率为0.01，即 $p=0.01$ 。取其中的 n 个，使得发生的个数除以总数的比例落在区间（0.0099,0.0101），则 n 至少是多少?

解：

首先，两点分布 B 的期望为 $\mu = p$ ，方差为 $\sigma^2 = \mathrm{p}(1-\mathrm{p})$ 。

其次，当 n 较大时，随机变量 $\mathrm{Y} = \sum_{i=1}^{n} B_i$ 近似服从正态分布，事实上，$\mathrm{X} = \frac{Y-n\mu}{\sqrt{n}\sigma} = \frac{\sum_{i=1}^{\mathrm{in}} B_i - n\mu}{\sqrt{n}\sigma}$ 近似服从标准正态分布。

从而:

$$P\left\{a \le \frac{\sum_{i=1}^{n} B_i}{n} \le b\right\} \ge 1-\alpha \Rightarrow P\left\{\frac{\sqrt{n}(a-\mu)}{\sigma} \le \frac{\sum_{i=1}^{n} B_i - n\mu}{\sigma} \le \frac{\sqrt{n}(b-\mu)}{\sigma}\right\} \ge 1-\alpha$$

$$\Rightarrow \Phi\left(\frac{\sqrt{n}(b-\mu)}{\sigma}\right) - \Phi\left(\frac{\sqrt{n}(a-\mu)}{\sigma}\right) \ge 1-\alpha$$

上式中，$\mu = 0.01$，$\sigma^2 = 0.0099$，$a = 0.0099$，$b = 0.0101$，$a = 0.05$或0.01（显著性水平的一般取值），查标准正态分布表，很容易计算得到 n 的最小值。

注：直接使用二项分布，也能得到结论。

中心极限定理的意义

- 实际问题中，很多随机现象可以看做许多因素的独立影响的综合反应，其往往近似服从正态分布。

例如：

城市耗电量：大量用户的耗电量总和

测量误差：许多观察不到的、微小的误差总和

值得注意的是：多个随机变量的和才可以，有些问题是乘性误差，则需要鉴别或者取对数后再使用。

- 线性回归中，将使用该定理论证最小二乘法的合理性。

  注：参考https://wenku.baidu.com/view/86b2a50e102de2bd9705886f.html

2.7.3 极大似然估计

极大似然估计是建立在极大似然原理的基础上的一个统计方法，而极大似然原理的直观想法是：一个随机试验如有若干个可能的结果 A，B，C，...，若在一次试验中，结果 A 出现了，那么可以认为实验条件对 A 的出现有利，也即出现的概率 $P(A)$ 较大。

一般说来，事件 A 发生的概率与某一未知参数 θ 有关， θ 取值不同，则事件 A 发生的概率 $P(A|\theta)$ 也不同，当我们在一次试验中事件 A 发生了，则认为此时的 θ 值应是 t 的一切可能取值中使达 $P(A|\theta)$ 到最大的那一个，极大似然估计法就是要选取这样的 t 值作为参数 t 的估计值，使所选取的样本在被选的总体中出现的可能性为最大。

举例：

设甲箱中有99个白球，1个黑球；乙箱中有1个白球，99个黑球。现随机取出一箱，再从抽取的一箱中随机取出一球，结果是黑球，这一黑球从乙箱抽取的概率比从甲箱抽取的概率大得多，这时我们自然更多地相信这个黑球是取自乙箱的。

- 求极大似然函数估计值的一般步骤：

  (1) 写出似然函数；

  (2) 对似然函数取对数，并整理；

  (3) 求导数；

  (4) 解似然方程 。

- 利用高等数学中求多元函数的极值的方法，有以下极大似然估计法的具体做法：

  (1) 根据总体的分布，建立似然函数 $L\left(x_1, x_2, \ldots, x_n ; \theta_1, \theta_2, \ldots, \theta_k\right)$；

  (2) 当 L 关于 θ_1，θ_2，$\ldots, \theta_k$ 可微时，(由微积分求极值的原理）可由方程组

$$\frac{\partial L}{\partial \theta_i}=0, i=1,2, \ldots, k$$

定出 $\hat{\theta}_i(i=1,2, \ldots, k)$，并称以上方程组为似然方程。

因为 L 与 lnL 有相同的极大值点，所以 $\hat{\theta}_i(i=1,2, \ldots, k)$ 也可由方程组

$$\frac{\partial lnL}{\partial \theta_i}=0, i=1,2, \ldots, k$$

定出 $\hat{\theta}_i(i=1,2, \ldots, k)$，并称以上方程组为对数似然方程；$\hat{\theta}_i(i=1,2, \ldots, k)$就是所求参数的 $\theta_i(i=1,2, \ldots, k)$ 极大似然估计量。

- 总体离散型与连续型极大似然估计

1.若总体 X 为离散型，其概率分布列为

$$P\{X=x\}=p(x ; \theta)$$

其中 θ 为未知参数。设 $\left(X_1, X_2, \ldots, X_n\right)$ 是取自总体的样本容量为 n 的样本，则 $\left(X_1, X_2, \ldots, X_n\right)$ 的联合分布律为 $\prod_{i=1}^n p\left(x_i ; \theta\right)$ 。又设 $\left(X_1, X_2, \ldots, X_n\right)$ 的一组观测值为 $\left(x_1, x_2, \ldots, x_n\right)$，易知样本 $X_1, X_2, \ldots, X_n$ 取到观测值的概率为 $L(\theta)=L\left(x_1, x_2, \ldots, x_n ; \theta\right)=\prod_{i=1}^n p\left(x_i ; \theta\right)$。

这一概率随 θ 的取值而变化，它是 θ 的函数，称 $L(\theta)$ 为样本的似然函数。

2.若总体 X 为连续型，其概率密度函数为 $f(x;\theta)$，其中 θ 为未知参数。设 $(X_1,X_2,\ldots,X_n)$ 是取自总体的样本容量为n的简单样本，则 $(X_1,X_2,\ldots,X_n)$ 的联合概率密度函数为 $\prod_{i=1}^{n} f(x_i;\theta)$ 。又设 $(X_1,X_2,\ldots,X_n)$ 的一组观测值为 $(x_1,x_2,\ldots,x_n)$，则随机点 $(X_1,X_2,\ldots,X_n)$ 落在点 $(x_1,x_2,\ldots,x_n)$ 的邻边（边长分别为$dx_1,dx_2,\ldots,dx_n$的 n 维立方体）内的概率近似地为 $\prod_{i=1}^{n} f(x_i;\theta)\,dx_i$。

考虑函数

$L(\theta)=L(x_1,x_2,\ldots,x_n;\theta)=\prod_{i=1}^{n} f(x_i;\theta)$。同样，$L(\theta)$ 称为样本的似然函数。

极大似然估计法原理就是固定样本观测值 $(x_1,x_2,\ldots,x_n)$，挑选参数 θ 使

$L\left(x_1,x_2,\ldots,x_n;\hat{\theta}\right)=\max L(x_1,x_2,\ldots,x_n;\theta)$。这样得到的 $\hat{\theta}$ 与样本值有关，$\hat{\theta}(x_1,x_2,\ldots,x_n)$称为参数 θ 的极大似然估计值，其相应的统计量称为 θ 的极大似然估计量。极大似然估计简记为MLE或 $\hat{\theta}$ 。

问题是如何把参数 θ 的极大似然估计 $\hat{\theta}$ 求出。更多场合是利用 $LnL(\theta)$ 是 $L(\theta)$ 的增函数，故 $LnL(\theta)$ 与在同一点处 $L(\theta)$ 达到最大值，于是对似然函数 $L(\theta)$ 取对数，利用微分学知识转化为求解对数似然方程

$\hat{\theta}(X_1,X_2,\ldots,X_n)\,\hat{\theta}(X_1,X_2,\ldots,X_n)\,\theta\frac{\partial \ln L(\theta)}{\partial \theta_j}=0, j=1,2,\ldots,k$，解此方程并对解做进一步的判断。但由最值原理，如果最值存在，此方程组求得的驻点即为所求的最值点，就可以很到参数的极大似然估计。极大似然估计法一般属于这种情况，所以可以直接按上述步骤求极大似然估计。

注：参考https://baike.sogou.com/v7678775.htm?fromTitle=%E6%9E%81%E5%A4%A7%E4%BC%BC%E7%84%B6%E4%BC%B0%E8%AE%A1