概率论基本概念

古典概型

计数原理 加法原理、乘法原理. 从含有 n 个元素的盒子中:

1. 有放回地取出 r 个元素组成的可重复排列的不同方式有 n^r 种;
2. 无放回地取出 r 个元素组成的不重复排列的不同方式有 $P_n^r = \dfrac{n!}{r!}$ 种（排列数），特别地，当 $r = n$ 时称为全排列;
3. 不放回地选取 r 个元素的组合有 $C_n^r = \dfrac{n!}{r!(n-r)!}$ 种（组合数）;
4. 有放回地选取 r 个元素的组合有 C_{n+r-1}^n 种（重复组合数）.

盒子模型 把 n 个球放到不同编号的 n 个盒子中:

1. 球可辨，每个盒子中不限球的个数，不同的放法个数为 n^r （重复排列）;
2. 球可辨，每个盒子中至多放一个球，不同的放法个数为 $\dfrac{n!}{(n-r)!}$ （选排列）;
3. 球不可辨，每个盒子中不限球的个数，不同的放法个数为 C_{n+r-1}^r （隔板法）;
4. 球不可辨，每个盒子中至多放一个球，不同的放法个数为 C_n^r .

多组组合 把 n 个不同的元祖分为有序的 k 个部分，第 i 部分有 r_i 个元素，不同的分法个数为 $\dfrac{n!}{r_1!r_2!\dots r_k!}$ （多项式系数）.

不尽相异元素的排列 有 n 个元素，属于 k 个不同的类，同类元素之间不可辨认，第 i 类元素有 n_i 个. 把这些元素排成一列，不同的排法个数为 $\dfrac{n!}{n_1!n_2!\dots n_k!}$.

概率的性质

1. $P(\varnothing) = 0$;
2. （两两不相容事件）有限可加性;
3. （子集事件）可减性;
4. （子集事件）单调性;
5. $P(\bar{A}) = 1 - P(A)$;
6. 容斥原理:

$$P\left(\bigcup_{k=1}^n A_k\right) = \sum_{m=1}^n (-1)^{m-1} \sum_{1\leqslant i<j\leqslant n} P(A_i \dots A_j)$$

其中第 m 项中概率函数的变量为 m 个事件的并.

7. 次可加性

$$P\left(\bigcup_{n=1}^\infty A_n\right) \leqslant \sum_{n=1}^\infty P(A_n)$$

8. 上连续性、下连续性.

条件概率

定义 $P(A|B) = \dfrac{P(AB)}{P(B)}$.

乘法公式 $P(AB) = P(A)P(B|A)$，归纳可得

$$P(A_1A_2\dots A_n) = P(A_1)P(A_2|A_1)\dots P(A_n|A_1A_2\dots A_n)$$

全概率公式 若 $\{B_n\}$ 是样本空间的一个完备事件群（无交、并为 Ω），则有

$$P(A) = \sum_{i=1}^n P(A|B_i)P(B_i)$$

Bayes 公式 若 $\{B_n\}$ 是样本空间的一个完备事件群（无交、并为 Ω），则有

$$P(B_i|A) = \frac{P(B_iA)}{P(A)} = \frac{P(A|B_i)P(B_i)}{\sum_{j=1}^n P(A|B_j)P(B_j)}$$

独立性

定义 两个事件满足 $P(AB) = P(A)P(B)$，则称 A, B 相互独立. 更一般地，

$$P\left(\bigcap_{i=1}^n A_i\right) = \prod_{i=1}^n P(A_i)$$

等价命题

1. $P(B|A) = P(B)$;
2. A 或 $\bar{A}$ 与 B 或 $\bar{B}$ 相互独立. 更一般地，令 $\tilde{A}_i = A_i$ 或 $\bar{A}_i$，有放回地取出

$$P\left(\bigcap_{i=1}^n \tilde{A}_i\right) = \prod_{i=1}^n P(\tilde{A}_i)$$

随机变量

离散型随机变量

概率密度函数 (pmf) 如果随机变量 X 只能取有限多个或可数多个值，那么称 X 为离散型随机变量，设 $\{x_k\}$ 为 X 所有取值的集合，则称

$$P(X = x_k) = p_k$$

为离散型随机变量 X 的分布律或**概率质量函数 (pmf)**.

无记忆性 以所有正整数为取值集合的随机变量 X 服从几何分布 $Ge(p)$，当仅当

$$P(X > m+n|X > m) = P(X > n)$$

这个性质被称为几何分布的无记忆性.

Poisson 逼近定理 设一族随机变量 $X_n \sim B(n, p_n)$，若当 $n \to \infty$ 时，$np_n \to \lambda > 0$，则有

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

实际应用中，$n \geqslant 30$, $np_n \leqslant 5$ 时即可应用.

连续型随机变量

累计分布函数 (cdf) 设 X 为一随机变量，$x \in \mathbb{R}$，称

$$F(x) = P(X \leqslant x)$$

为随机变量 X 的（累计）分布函数.

概率密度函数 (pdf) 若存在可积的非负函数 $f(x) \geqslant 0$，使得

$$\forall x \in \mathbb{R}, \quad F(x) = \int_{-\infty}^x f(t)\mathrm{d}t$$

则称 X 为连续型随机变量，$f(x)$ 称为分布函数或**概率密度函数 (pdf)**，记为 $X \sim f(x)$.

概率密度函数的性质

1. 恒为非负;
2. $\displaystyle\int_{-\infty}^{+\infty} f(x)\mathrm{d}x = 1$;
3. 对任意可测集合 $A \subseteq \mathbb{R}$，有

$$P(X \in A) = \int_A f(x)\mathrm{d}x$$

4. 若 $f(x)$ 在 x_0 连续，则有 $F'(x_0) = f(x_0)$;
5. $\forall x \in \mathbb{R}$, $P(X = x) = 0$

正态分布的性质 主要针对其概率密度函数的性质:

1. “钟型”曲线，两头小，中间大，关于 $x = \mu$ 对称;
2. 最大值在 $x = \mu$ 处取得 $f(\mu) = \dfrac{1}{\sqrt{2\pi}\sigma}$;
3. $x = \mu \pm \sigma$ 为拐点，图形以 x 轴为渐进线;
4. μ 决定图形位置，σ 决定图形形状.

标准正态分布 标准正态分布 $X \sim N(0,1)$ 的概率密度函数记为 $\phi(x)$，累计分布函数记为 $\Phi(x)$，显然有 $\Phi(-x) = 1 - \Phi(x)$.

正态分布标准化 对于正态分布 $X \sim N(\mu, \sigma^2)$，有下式恒成立

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

随机变量函数的分布 设 $X \sim f(x)$，随机变量 $Y = g(X)$，则有

$$F_Y(y) = \int_{g(x)\leqslant y} f(x)\mathrm{d}x$$

若 $g(x)$ 在各个区间 I_i 上严格单调且反函数可导时，设 $X = h(Y)$，则有

$$f_Y(y) = \sum_i f(h(y))|h'(y)|I_i(y)$$

多维随机变量

联合分布函数 (joint cdf) 设 (X, Y) 是二维随机变量，称二元函数

$$F(x, y) = P(X \leqslant, x, Y \leqslant y)$$

为 (X, Y) 的**联合分布函数**.

联合概率密度函数 (joint pdf) 设 $(X, Y) \sim F(x, y)$ ，若存在可积的非负函数 $f(x, y) \leqslant 0$，使得

$$\forall (x, y) \in \mathbb{R}^2, \quad F(x, y) = \int_{-\infty}^x \int_{-\infty}^y f(u, v)\mathrm{d}u\mathrm{d}v$$

则称 (X, Y) 为而为连续型随机变量，$f(x, y)$ 称为**联合概率密度函数**.

边缘分布 设 $(X,Y)\sim F(x,y)$，$X\sim F_X(x), Y\sim F_Y(y)$，则称 $F_X(x)$ 和 $F_Y(y)$ 为 (X,Y) 或 F 的边际分布 (marginal distribution). 对应的 $f_X(x)$ 和 $f_Y(y)$ 称为 $f(x,y)$ 的边际概率密度函数 (marginal pdf). 且有

$$F_X(x)=\lim_{y\to\infty}F(x,y),\quad f_X(x)=\int_{-\infty}^{+\infty}f(x,y)\mathrm{d}y$$

条件概率密度函数 (conditional pdf) 给定 $Y=y$ 条件下随机变量 X 的条件概率密度函数为

$$f_{X|Y}(x|y)=\frac{f(x,y)}{f_Y(y)}$$

n 维情况 对于一组随机变量 $(X_1,X_2,\dots,X_n)\sim F$，记 $\boldsymbol{U}=\{X_1,X_2,\dots,X_k\}$，$\boldsymbol{V}=\{X_{k+1},X_{k+2},\dots,X_n\}$，则 $\boldsymbol{U}$ 的边缘分布为

$$F_{\boldsymbol{U}}(\boldsymbol{u})=F(x_1,x_2,\dots,x_k,\infty,\infty,\dots,\infty)$$

边缘密度函数为

$$f_{\boldsymbol{U}}(\boldsymbol{u})=\int_{\mathbb{R}^{n-k}}f(\boldsymbol{u},\boldsymbol{v})\mathrm{d}\boldsymbol{v}$$

n 维随机变量的边际分布函数有 $n-2$ 个. 给定 $\boldsymbol{V}=\boldsymbol{v}$ 条件下随机变量 $\boldsymbol{U}$ 的条件概率密度函数为

$$f_{\boldsymbol{U}|\boldsymbol{V}}(\boldsymbol{u}|\boldsymbol{v})=\frac{f(\boldsymbol{u},\boldsymbol{v})}{f_{\boldsymbol{V}}(\boldsymbol{v})}$$

随机变量独立性 若 $(X,Y=)\sim F(x,y)$, $X\sim F_X(x)$, $Y\sim F_Y(y)$，若

$$\forall(x,y)\in\mathbb{R}^2,\quad F(x,y)=F_X(x)F_Y(y)$$

则称随机变量 X,Y 相互独立. 可以推广到 n 维情况.

随机向量函数的分布 一维情况 $Z=g(X,Y)$，则有

$$F_Z(z)=\iint_{g(x,y)\leqslant z}f(x,y)\mathrm{d}x\mathrm{d}y$$

二维情况 $(Z_1,Z_2)=(g_1(X,Y),g_2(X,Y))$，则有

$$F_{\boldsymbol{Z}}(z_1,z_2)=\iint_{g_1(x,y)\leqslant z_1\,g_2(x,y)\leqslant z_2}f(x,y)\mathrm{d}x\mathrm{d}y$$

若 g_1,g_2 是一一映射，$(x,y)=(h_1(z_1,z_2),h_2(z_1,z_2))$，则有

$$f_{\boldsymbol{Z}}(z_1,z_2)=f(h_1(z_1,z_2),h_2(z_1,z_2))\left|\frac{\partial(h_1,h_2)}{\partial(u,v)}\right|_{(u,v)=(z_1,z_2)}$$

更一般的，若对于 n 维随机变量 $\boldsymbol{X}\sim f_{\boldsymbol{X}}(\boldsymbol{x})$，存在 n 维随机变量函数 $\boldsymbol{Y}=\boldsymbol{g}(\boldsymbol{X})$，且 $\boldsymbol{g}:\mathbb{R}^n\to\mathbb{R}^n$ 是一一映射，则

$$f_{\boldsymbol{Y}}(\boldsymbol{y})=f_{\boldsymbol{X}}(\boldsymbol{g}^{-1}(\boldsymbol{y}))|\boldsymbol{J}|I_D(y)$$

其中 $D\subset\mathbb{R}^n$ 是 $\boldsymbol{y}$ 的密度非零的所有取值的集合，$\boldsymbol{J}$ 是变换 $\boldsymbol{g}^{-1}$ 对 $\boldsymbol{y}$ 的 Jacobi 矩阵.

随机变量数字特征

期望 离散型随机变量 X 的期望

$$E(X)=\sum_{k\geqslant1}x_kp_k\leqslant\infty$$

连续型随机变量 $X\sim f(x)$ 的期望

$$E(X)=\int_{\mathbb{R}}xf(x)\leqslant\infty$$

期望的性质

1. 线性性
2. 对于相互独立随机变量有:

$$E(X_1X_2\dots X_n)=\prod_{i=1}^{n}E(X_i)$$

3. 对于 $\boldsymbol{X}\sim f(\boldsymbol{x}),\boldsymbol{Y}=g(\boldsymbol{X})$，有（离散同理）:

$$E(\boldsymbol{Y})=\int_{\mathbb{R}^n}g(\boldsymbol{x})f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$$

马尔科夫不等式 若随机变量 $X\geqslant0$，则

$$\forall\varepsilon>0,\quad P(X\geqslant\varepsilon)\leqslant\frac{E(X)}{\varepsilon}$$

条件期望 称下式为给定 $X=x$ 时随机变量 Y 的条件期望

$$E(Y|X=x)=\int_{\mathbb{R}}yf_{Y|X}(y|x)\mathrm{d}y$$

$E(Y|X)$ 是关于 X 的随机变量，且满足**条件期望的平滑公式(全期望公式)**:

$$E(E(Y|X))=E(Y)$$

中位数 随机变量 $X\sim F(x)$，**中位数** m 满足:

$$P(X\geqslant m)=1-F(m-0)\geqslant\frac{1}{2},\quad P(X\leqslant m)=F(m)\geqslant\frac{1}{2}$$

众数 使得随机变量 X 的概率质量函数 (pmf, X 离散时) 或概率密度函数 (pdf，X 连续时) 达到最大的常数 m_d 称为**众数**.

p 分位数 设 $p\in(0,1)$，随机变量 X 的 p 分位数 Q_p 定义为

$$P(X\leqslant Q_p)\geqslant p,\quad P(X\geqslant Q_p)\geqslant1-p$$

定义内四分距 $IQR=Q_{0.75}-Q_{0.25}$.

矩 随机变量 X，满足 $E(|X|^k)<\infty$，则称 $E(X-c)^k$ 为 X 关于 c 的 k 阶矩；称 $\alpha_k=E(X^k)$ 为 X 的 k 阶原点矩；称 $\mu_k=E(X-E(X^k))$ 为 X 的 k 阶中心距.

矩母函数 随机变量 X 的矩母函数 (MGF) 定义为

$$M_X(s)=E\mathrm{e}^{sX}$$

矩母函数唯一决定随机变量分布.

方差和标准差 随机变量 X 关于其均值 μ 的二阶矩称为 X 的**方差**

$$\sigma^2=Var(X)=E((X-\mu)^2)$$

方差的算术平方根称为 X 的**标准差**

$$\sigma=\sqrt{\sigma^2}$$

方差的性质

1. $Var(X)=E(X^2)-\mu^2$;
2. $Var(cX)=c^2Var(X)$, $Var(X+d)=Var(X)$;
3. 独立随机变量和的方差等于方差的和;
4. $Var(X)\leqslant E(X-c)^2,\quad \text{when } c=E(X)$.

协方差 随机变量 X 或 Y 均平方可积且平方均值有限，则随机变量 X,Y 的**协方差**为

$$Cov(X,Y)=E[(X-E(X))(Y-E(Y))]$$

协方差的性质

1. $Cov(X,Y)=Cov(Y,X)$;
2. $Cov(X,Y)=E(XY)-E(X)E(Y)$
3. $$Cov(aX+bY,cX+dY)=$$
$$(a,b)\begin{pmatrix}Var(X)&Cov(X,Y)\\Cov(X,Y)&Var(Y)\end{pmatrix}\begin{pmatrix}c\\d\end{pmatrix}$$
4. 随机变量的 Cauchy-Schwarz 公式:

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

相关系数 随机变量 X,Y 协方差存在，定义其**相关系数**为

$$\rho_{X,Y}=\frac{(X,Y)}{\sigma_X\sigma_Y}$$

独立与不相关 若随机变量相互独立，则随机变量间的相关系数为 0；反之不必成立，但在随机变量服从二元正态分布时二者等价.

对于任何非退化随机变量 X,Y，如下四个命题等价;

1. X 与 Y 不相关;
2. $Cov(X,Y)=0$;
3. $E(XY)=E(X)E(Y)$;
4. $Var(X+Y)=Var(X)+Var(Y)$.

熵 离散型随机变量的**熵**定义为

$$H(X)=-\sum_{k=1}^{\infty}p_k\log_2(p_k)$$

连续型随机变量的熵定义为

$$H(X)=-\int_{-\infty}^{+\infty}f_X(x)\ln f_X(x)\mathrm{d}x$$

极限定理

对于 i.i.d. 随机变量 $X_1,X_2,\dots,X_n$，记 $S_n=\sum_{i=1}^{n}X_i$，则有

大数定律 $\forall\varepsilon>0,\quad\lim_{n\to\infty}P\left(\left|\frac{S_n}{n}-\mu\right|\geqslant\varepsilon\right)=0$

Lindeberg-Levy 中心极限定理 $\left(\frac{\sqrt{n}(S_n/n-\mu)}{\sigma}\right)\sim N(0,1)$

DeMoivre-Laplace 定理 $X_i\sim B(1,p),\quad\frac{S_n/n-np}{\sqrt{np(1-p)}}\sim N(0,1)$

统计学基本概念

样本 总体中按一定方式抽取的 n 个个体被称为是样本量为 n 个一个样本，记为 $\boldsymbol{X}=(X_1,X_2,\dots,X_n)$.

放回抽样得到的样本是**简单随机样本**，简单随机样本中 $X_1,X_2,\dots,X_n$ 满足 i.i.d. 简单样本的联合分布函数为 $\prod_{i=1}^{n}F(x_i)$，联合概率密度函数（若存在）为 $\prod_{i=1}^{n}f(x_i)$.

常见统计量 完全由样本 $\boldsymbol{X}$ 决定（不含有参数）的量被称为统计量，常见统计量有

1. 样本均值：$\bar{X}=\dfrac{1}{n}\sum_{i=1}^{n}X_i$
2. 样本方差：$S^2=\dfrac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$，其中 $S=\sqrt{S^2}$ 被称为样本标准差
3. 样本 k 阶原点矩：$a_k=\dfrac{1}{n}\sum_{i=1}^{n}X_i^k$
4. 样本 k 阶中心矩：$m_k=\dfrac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^k$
5. 样本偏度系数：$\hat{\beta}_1=\dfrac{m_3}{m_2^{3/2}}$
6. 样本峰度系数：$\hat{\beta}_2=\dfrac{m_4}{m_2^2}$
7. 样本相关系数：

$$\rho_n=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\left(\sum_{i=1}^{n}(X_i-\bar{X})^2\sum_{i=1}^{n}(Y_i-\bar{Y})^2\right)}}$$

8. 次序统计量：把样本 $\boldsymbol{X}$ 从小到大排列为

$$X_{(1)}\leqslant X_{(2)}\leqslant\cdots\leqslant X_{(n)}$$

9. 样本中位数：

$$m_n=\begin{cases}X_{(\frac{n+1}{2})}, & \text{when } n \text{ is odd}\\ \dfrac{1}{2}\left(X_{(\frac{n}{2})}+X_{(\frac{n}{2}+1)}\right), & \text{when } n \text{ is even}\end{cases}$$

10. 经验分布函数 $F_n(x)=\dfrac{N(x)}{n}$，其中 $N(x)$ 是 $\boldsymbol{X}$ 中的样本数据小于等于 x 的个数

χ^2 分布 $\boldsymbol{X}$ 是来自标准正态分布总体的一个简单随机样本，称

$$X:=\sum_{i=1}^{n}X^2$$

服从自由度为 n 的 χ^2 分布，记为 $X\sim\chi_n^2$. 显然有 χ_n^2 分布的概率密度函数

$$k_n(x)=\frac{1}{\Gamma(\frac{n}{2})2^{\frac{n}{2}}}\mathrm{e}^{-\frac{x}{2}}x^{\frac{n-2}{2}}I_{(0,+\infty)}(x)$$

χ^2 分布的性质：

1. 若 $X\sim\chi_n^2$，则有 $E(X)=n, Var(X)=2n$；
2. 若 $X\sim\chi_m^2, Y\sim\chi_n^2$ 且 X,Y 相互独立，则 $X+Y\sim\chi_{m+n}^2$.

t 分布 设 $X\sim N(0,1), Y\sim\chi_n^2$，且 X,Y 相互独立，称

$$T=\frac{X}{\sqrt{Y/n}}$$

服从自由度为 n 的 t 分布，记为 $T\sim t_n$. 显然有 t_n 分布的概率密度函数

$$f_n(t)=\frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}},\quad t\in\mathbb{R}$$

t 分布的性质：

1. 当 $n=1$ 时，t 分布就是 Cauchy 分布；
2. 若 $T\sim t$，当 $n\geqslant 2$ 时，$E(T)=0$；当 $n\geqslant 3$ 时 $Var(T)=\dfrac{n}{n-2}$；
3. $\lim\limits_{n\to\infty}f_n(t)=\varphi(t)$，其中 $\varphi(t)$ 是标准正态分布的概率密度函数.

F 分布 设 $X\sim\chi_m^2, Y\sim\chi_n^2$，且 X,Y 相互独立，称

$$F=\frac{X/m}{Y/n}$$

服从自由度为 m,n 的 F 分布，记为 $F\sim F_{m,n}$. 显然有 $F_{m,n}$ 分布的概率密度函数

$$f_{m,n}(x)=m^{\frac{n}{2}}n^{\frac{n}{2}}\frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})}x^{\frac{m}{2}-1}(mx+n)^{-\frac{m+n}{2}}I_{(0,+\infty)}(x)$$

F 分布的性质：

1. 若 $Z\sim F_{m,n}$，则 $\dfrac{1}{Z}\sim F_{m,n}$；
2. 若 $T\sim t_n$，则 $T^2\sim F_{1,n}$；
3. $F_{m,n}(1-\alpha)=\dfrac{1}{F_{n,m}(\alpha)}$

正态分布的样本 设简单样本 $\boldsymbol{X}$ 均来自 $N(\mu,\sigma)$ 的总体，则有

1. 线性统计量满足参正态分布：

$$T=\sum_{i=1}^{n}c_iX_i\sim N(\mu\sum_{i=1}^{n}c_i,\sigma^2\sum_{i=1}^{n}c_i^2)$$

2. 样本均值 $\bar{X}\sim N(\mu,\dfrac{\sigma^2}{n})$
3. 样本方差 $\dfrac{(n-1)S^2}{\sigma^2}\sim\chi_{n-1}^2$
4. $\bar{X}$ 与 S^2 相互独立，故 $\dfrac{\sqrt{n}(\bar{X-}\mu)}{S}\sim t_{n-1}$

两个立刻的推论是若 m 样本量 $\boldsymbol{X}$ 和 n 样本量样本 $\boldsymbol{Y}$ 所有变量相互独立，且分别来自样本 $N(\mu_1,\sigma_1^2)$, $N(\mu_2,\sigma_2^2)$，则有

1.

$$T=\frac{(\bar{X}-\bar{Y}-(\mu_1-\mu_2))}{S_T}\sqrt{\frac{mn}{m+n}}\sim t_{m+n-2}$$

其中 S_T 满足 $(m+n-2)S_T^2=(m-1)S_X^2+(n-1)S_Y^2$

2.

$$F=\frac{S_X^2}{S_Y^2}\cdot\frac{\sigma_2^2}{\sigma_1^2}\sim F_{m-1,n-1}$$

指数分布的样本 若 $\boldsymbol{X}$ 来自参数为 λ 的指数分布则有

$$2\lambda n\bar{X}=2X\sum_{i=1}^{n}X_i\sim\chi_{2n}^2$$

参数点估计

点估计 总体分布中的参数 θ_k（例如正态分布 $N(\mu,\sigma^2)$ 中的 μ,σ 可以用统计量 $\hat{\theta}(\boldsymbol{X})$ 去估计，用数轴上一个点 $\hat{\theta}(\boldsymbol{x})$ 去估计一个点 θ 的估计称为**点估计**.

矩估计 用样本矩估计总体矩. 例如用一阶原点矩估计总体期望，用二阶中心矩估计总体方差等.

最大似然估计 若 $\boldsymbol{X}$ 有联合概率密度函数 $f(\boldsymbol{x};\theta)$，其中 $\theta=(\theta_1,\theta_2,\dots,\theta_n)$，固定 $\boldsymbol{x}$ 时，称关于 θ 的函数 $L(\theta;\boldsymbol{x})=f(\boldsymbol{x};\theta)$ 为 θ 的似然函数.

当参数 $\theta=\theta^\star$ 时 $L(\theta,\boldsymbol{x})$ 取最大值，则可将 $\theta^\star$ 作为 θ 的一个估计值，称为最大似然估计.

若似然函数光滑，且样本为简单随机样本，则可取 $\ell=\ln L$（称为对数似然函数）来化简求解过程.

偏差 $\hat{g}(\boldsymbol{X})$ 是 $g(\theta)$ 的一个估计量，则偏差为

$$E_\theta(\hat{g}(\boldsymbol{X}))-g(\theta)$$

偏差为 0 的估计为无偏估计.

样本数据的加权和是总体期望的无偏估计. 样本方差是总体方差的无偏估计，二阶中心矩估计不是.

最小方差无偏估计 (MVUE) 均方误差

$$MSE_\theta(\hat{\theta})=E_\theta(\hat{\theta}(\boldsymbol{X})-\theta)^2$$

平均绝对误差

$$MAE_\theta(\hat{\theta})=E_\theta(|\hat{\theta}(\boldsymbol{X})-\theta|)$$

无偏估计下 $MSE_\theta(\hat{\theta})=Var_\theta(\hat{\theta})$. 若 $Var_\theta(\hat{\theta}_1)\leqslant Var_\theta(\hat{\theta}_2),\forall\theta$，且存在 θ 使等号成立，则称 $\hat{\theta}_1$ 更有效. 对于任意无偏估计 $\hat{\theta}$ 满足 $Var_\theta(\hat{\theta}^\star)\leqslant Var(\hat{\theta})$ 的 $\hat{\theta}^\star$ 被称为最小方差无偏估计，即在均方误差标准下最有效的估计.

样本数据的加权和作为总体期望的估计时，均值（权为 $\dfrac{1}{n}$ 的估计最有效.

克拉默-拉奥方差下界 对于 $g(\theta)$ 的无偏估计 $\hat{g}(\boldsymbol{X})$，在正则条件下有

$$Var_\theta(\hat{g}(\boldsymbol{X}))\geqslant(g'(\theta))^2[nI(\theta)]^{-1}$$

其中 $I(\theta)=E\left(\dfrac{\partial\ln f(X;\boldsymbol{\theta})}{\partial\theta}\right)^2$ 被称为费希尔信息函数。

$\bar{X}$ 是 μ 的 MUVE，当 μ 已知时，$\dfrac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2$ 是 σ^2 的一个 MUVE.

相合性 若点估计 $\hat{\theta}$ 在样本量 $n\to\infty$ 时依概率趋近于 θ，则称 $\hat{\theta}$ 是 θ 的一个（弱）相合估计量.

相合性是对一个估计量的最基本要求. 若一个估计量没有相合性，则无论样本量多大也不能把未知参数估计到任意精度，这种估计是不可取的.

渐进正态性 若点估计 $\hat{\theta}$ 在样本量 $n\to\infty$ 时

$$\left(\frac{\hat{\theta}(\boldsymbol{X})-\theta}{Var_\theta(\hat{\theta}(\boldsymbol{X}))}\sim N(0,1)\right)$$

则称 $\hat{\theta}$ 有渐进正态性.

区间估计

置信区间 参数 θ，统计量 $\hat{\theta}_1, \hat{\theta}_2$，则参数 θ 的置信系数为 $1-\alpha$ 的区间是 $[\hat{\theta}_1, \hat{\theta}_2]$ 可表示为

$$P(\hat{\theta}_1 \leqslant \theta \leqslant \hat{\theta}_2) = 1-\alpha$$

α 是一个小的正数，通常取 $0.01, 0.05, 0.1$.

枢轴变量法 可以分为以下步骤：

1. 找一个 θ 的良好点估计 $T(\boldsymbol{X})$，一般为最大似然估计（例如 $T(\boldsymbol{X}) = \bar{X}$ 估计 $\theta = \mu$）；
2. 构造枢轴 $S(T, U, \theta)$ 使得 S 分布已知（设其概率分布函数为 F），其中 U 为统计量（例如构造 $\dfrac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \sim N(0,1)$）；
3. 枢轴变量应满足 $w_{1-\alpha} \leqslant S \leqslant w_\alpha \Longrightarrow \hat{\theta}_1 \leqslant \theta \leqslant \hat{\theta}_2$，其中 w_α 为 F 的上 α 分位数（满足 $F(w_\alpha) = 1-\alpha$）.

正态总体均值 μ 的置信区间 设置信区间为 $\bar{x} \pm d$，则误差界限 d 为（用 $\hat{\sigma}$ 估计时总体不必为正态分布）

$$d = \begin{cases} \dfrac{\sigma}{\sqrt{n}} u_{\frac{\alpha}{2}}, & \text{when } \sigma^2 \text{ is known} \\ \dfrac{s}{\sqrt{n}} t_{n-1}\left(\dfrac{\alpha}{2}\right), & \text{when } \sigma^2 \text{ is unknown} \\ \dfrac{\hat{\sigma}}{\sqrt{n}} u_{\frac{\alpha}{2}}, & \text{when } n > 30 \text{ and } \sigma^2 \text{ is unknown} \end{cases}$$

正态总体方差 σ 的置信区间 总体均值未知，可考虑枢轴估计 $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$

$$\sigma^2 \in \left[\frac{(n-1)s^2}{\chi^2_{n-1}(\frac{\alpha}{2})}, \frac{(n-1)s^2}{\chi^2_{n-1}(1-\frac{\alpha}{2})}\right]$$

两个正态总体均值差 $\mu_2 - \mu_1$ 的置信区间 设置信区间为 $\bar{y} - \bar{x} \pm d$，则误差界限 d 为

$$d = \begin{cases} \sqrt{\dfrac{\sigma_1^2}{m} + \dfrac{\sigma_2^2}{n}} u_{\frac{\alpha}{2}}, & \text{when } \sigma_1^2, \sigma_2^2 \text{ are knonw} \\ \sqrt{\dfrac{m+n}{mn}} s_T t_{m+n-2}\left(\dfrac{\alpha}{2}\right), & \text{when } \sigma_1^2, \sigma_2^2 \text{ are unknonw} \end{cases}$$

两个正态总体方差比 $\dfrac{\sigma_1^2}{\sigma_2^2}$ 的置信区间 总体均值未知，可考虑枢轴估计 $\dfrac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{m-1,n-1}$

$$\frac{\sigma_1^2}{\sigma_2^2} \in \left[\frac{s_1^2}{s_2^2} F_{n-1,m-1}\left(1-\frac{\alpha}{2}\right), \frac{s_1^2}{s_2^2} F_{n-1,m-1}\left(\frac{\alpha}{2}\right)\right]$$

比例 p 的区间估计 事件 A 在每次试验中发生概率为 p，若在 n 次试验中发生了 y_n 次，则 p 的 $1-\alpha$ 置信区间为

$$p \in \frac{\hat{p}+\delta}{1+2\delta} \pm \frac{u_{\frac{\alpha}{2}}}{\sqrt{n}} \frac{\sqrt{\hat{p}(1-\hat{p}) + \delta/2}}{1+2\delta}$$

其中 $\hat{p} = \dfrac{y_n}{n}$，$\delta = \dfrac{u_{\frac{\alpha}{2}}^2}{2n}$.

一般来说估计 p 置信区间需要 $n > 100$，在实际中我们可以忽略 δ，即 $p \in \hat{p} \pm \dfrac{u_{\frac{\alpha}{2}}}{\sqrt{n}} \sqrt{\hat{p}(1-\hat{p})}$.

置信区间宽度 $w = \dfrac{u_{\frac{\alpha}{2}}}{\sqrt{n}} \dfrac{\sqrt{\hat{p}(1-\hat{p}) + \delta/2}}{1+2\delta}$ 时可以计算要求的样本量 $n = \dfrac{4u_{\frac{\alpha}{2}}^2 \hat{p}(1-\hat{p})}{w^2}$.

基本自助置信区间 在总体中抽取样本 $\{x_n\}$，从样本中有放回的抽取一组样本量同样为 n 的样本 $\{x_n^\star\}$，称为一个自助样本，基于自助样本计算统计量 $\hat{\theta}$ 的值，称为 $\hat{\theta}$ 的一个自助版本，重复 B 次，在 B 个自助版本中记 $\dfrac{\alpha}{2}$ 分位数为 a，$1-\dfrac{\alpha}{2}$ 分位数为 b，则 θ 的 $1-\alpha$ 基本自助区间为

$$[2\hat{\theta} - b, 2\hat{\theta} - a]$$

自助 t 置信区间 记自助版本 $\hat{\theta}^\star$ 的标准差估计为

$$\hat{se}_B(\hat{\theta}_\star) = \sqrt{\frac{1}{B-1}\sum_{i=1}^{B}(\hat{\theta}_i^\star - \bar{\hat{\theta}}^\star)^2}$$

第 b 个重复下的“T 类型”为

$$t^{(b)} = \frac{\hat{\theta}_b^\star = \hat{\theta}}{\hat{se}_R(\hat{\theta}_b^\star)}$$

其中 $\hat{se}_R(\hat{\theta}_b^\star)$ 是对第 b 个自助样本再采用自助法生成 R 个自助版本得到的 $\hat{\theta}_b^\star$ 的标准差.

记“T 类型”中的则最终得到的 $\dfrac{\alpha}{2}$ 分位数为 t_a，$1-\dfrac{\alpha}{2}$ 分位数为 t_b θ 的 $1-\alpha$ 自助 t 置信区间为

$$[\hat{\theta} - t_b \hat{se}_B(\hat{\theta}), \hat{\theta} - t_a \hat{se}_B(\hat{\theta})]$$

置信限 参数 θ，统计量 $\overline{\theta}, \underline{\theta}$，则参数 θ 的置信系数为 $1-\alpha$ 的置信上限是 $\overline{\theta}$ 可表示为

$$P(\overline{\theta} \geqslant \theta) = 1-\alpha$$

参数 θ 的置信系数为 $1-\alpha$ 的置信下限是 $\underline{\theta}$ 可表示为

$$P(\theta \geqslant \underline{\theta}) = 1-\alpha$$

例如正态总体均值 μ 的置信上限在 σ^2 已知时可以写为 $\bar{x} + \dfrac{\sigma}{\sqrt{n}} u_\alpha$.

假设检验

假设的基本概念 原假设 H_0、备择假设 H_1，常见的假设：两点假设、双边假设、单边假设；检验一个假设时用到的统计量称为假设统计量，使假设得到接受的样本集合（样本所在区域）为接受域，被拒绝的区域为拒绝域，接受域和拒绝域的边界为临界值.

功效函数 功效函数：根据样本 $\boldsymbol{X}$ 所做的一个检验 Ψ，对应的功效函数为

$$\beta_\Psi(\theta) = P_\theta(\text{在检验 } \Psi \text{ 下假设 } H_0 \text{ 被否定})$$

根据功效函数可以得到检验 Ψ 的检验水平 α

$$\beta_\Psi(\theta) \leqslant \alpha, \quad \forall \theta in H_0$$

两类错误

1. 事实上 H_0 成立时，若检验 Ψ 拒绝了 H_0，则称为第一类错误，即“弃真错误“，弃真错误发生的概率为 $\alpha_{1\Psi}(\theta) = \beta_\Psi(\theta), \theta \in H_0$；
2. 事实上 H_0 不成立时，若检验 Ψ 接受了 H_0，则称为第二类错误，即“存伪错误“，存伪错误发生的概率为 $\alpha_{2\Psi}(\theta) = 1 - \beta_\Psi(\theta), \theta \in H_1$；

显著性检验 仅考虑第一类错误的检验称为显著性检验，显著性检验的一般方法如下：

1. 求出未知参数 θ 的一个较优的点估计 $\hat{\theta}$；
2. 寻找一个检验统计量 $T(\hat{\theta}, U, \theta_0)$，使得 $\theta = \theta_0$ 时 T 的分布已知；
3. 根据备择假设的实际意义寻找 T 的拒绝域；
4. 计算在原假设成立的条件下犯第一类错误的小于等于给定的显著性水平 α，得到一个临界值（T 在 $\theta = \theta_0$ 的 α 分位数，确定拒绝域；
5. 根据已有数据计算检验统计量是否在拒绝域中；
6. 根据具体问题解释是否能拒绝原假设.

单个正态总体均值的检验 设定假设

(1) $H_0 : \mu \geqslant \mu_0 \leftrightarrow H_1 : \mu < \mu_0$

(2) $H_0' : \mu \leqslant \mu_0 \leftrightarrow H_1' : \mu > \mu_0$

(3) $H_0'' : \mu = \mu_0 \leftrightarrow H_1'' : \mu \neq \mu_0$

当 σ^2 已知时，考虑统计量 $Z = \dfrac{\sqrt{n}(\bar{X}-\mu_0)}{\sigma} \sim N(0,1)$，水平为 α 的检验为

1. Ψ：当 $Z < -u_\alpha$ 时拒绝 H_0，否则不能拒绝 H_0；
2. Ψ'：当 $Z > u_\alpha$ 时拒绝 H_0'，否则不能拒绝 H_0'；
3. Ψ''：当 $|Z| > u_{\frac{\alpha}{2}}$ 时拒绝 H_0''，否则不能拒绝 H_0''.

当 σ^2 未知时，考虑统计量 $T = \dfrac{\sqrt{n}(\bar{X}-\mu_0)}{S} \sim t_{n-1}$，水平为 α 的检验为

1. Φ：当 $T < -t_{n-1}(\alpha)$ 时拒绝 H_0，否则不能拒绝 H_0；
2. Φ：当 $T > t_{n-1}(\alpha)$ 时拒绝 H_0'，否则不能拒绝 H_0'；
3. Φ：当 $|T| > t_{n-1}\left(\dfrac{\alpha}{2}\right)$ 时拒绝 H_0''，否则不能拒绝 H_0''.

设立原假设和备择假设的两条原则

1. 把已有的经过考验的结论或事实作为原假设 H_0；
2. 把希望得到的结论放在备择假设 H_1，希望能通过拒绝原假设得到希望得到的结论.

成组比较两个正态总体均值差 设定假设

(1) $H_0 : \mu_1 - \mu_2 \geqslant \delta \leftrightarrow H_1 : \mu_1 - \mu_2 < \delta$

(2) $H_0' : \mu_1 - \mu_2 \leqslant \delta \leftrightarrow H_1' : \mu_1 - \mu_2 > \delta$

(3) $H_0'' : \mu_1 - \mu_2 = \delta \leftrightarrow H_1'' : \mu_1 - \mu_2 \neq \delta$

当 σ^2 已知时，考虑统计量 $Z = \dfrac{\bar{X} - \bar{Y} - \delta}{\sigma\sqrt{\frac{1}{m} + \frac{1}{n}}} \sim N(0,1)$，水平为 α 的检验为

1. g：当 $Z < -u_\alpha$ 时拒绝 H_0，否则不能拒绝 H_0；
2. g'：当 $Z > u_\alpha$ 时拒绝 H_0'，否则不能拒绝 H_0'；
3. g''：当 $|Z| > u_{\frac{\alpha}{2}}$ 时拒绝 H_0''，否则不能拒绝 H_0''.

当 σ^2 未知时，考虑统计量 $T = \sqrt{\dfrac{mn}{m+n}} \dfrac{\bar{X} - \bar{Y} - \delta}{S_T} \sim t_{m+n-2}$，水平为 α 的检验为

1. h：当 $T < -t_{m+n-2}(\alpha)$ 时拒绝 H_0，否则不能拒绝 H_0；
2. h'：当 $T > t_{m+n-2}(\alpha)$ 时拒绝 H_0'，否则不能拒绝 H_0'；
3. h''：当 $|T| > t_{m+n-2}\left(\dfrac{\alpha}{2}\right)$ 时拒绝 H_0''，否则不能拒绝 H_0''.

成对比较两个正态总体均值差 构造新的样本 $Z_i = X_i - Y_i$，并对 $\boldsymbol{Z}$ 的均值做假设检验.

正态总体方差的检验 设定假设

(1) $H_0 : \sigma^2 \geqslant \sigma_0^2 \leftrightarrow H_1 : \sigma^2 < \sigma_0^2$

(2) $H_0' : \sigma^2 \leqslant \sigma_0^2 \leftrightarrow H_1' : \sigma^2 > \sigma_0^2$

(3) $H_0'' : \sigma^2 = \sigma_0^2 \leftrightarrow H_1'' : \sigma^2 \neq \sigma_0^2$

考虑统计量 $\chi^2 = \dfrac{(n-1)S^2}{\sigma_0^2} \sim \chi_{n-1}^2$，水平为 α 的检验为

1. ϕ：当 $\chi^2 < \chi_{n-1}^2(1-\alpha)$ 时拒绝 H_0，否则不能拒绝 H_0；
2. ϕ'：当 $\chi^2 > \chi_{n-1}^2(\alpha)$ 时拒绝 H_0'，否则不能拒绝 H_0'；
3. ϕ''：当 $\chi^2 < \chi_{n-1}^2\left(1-\dfrac{\alpha}{2}\right)$ 时拒绝 H_0''，否则不能拒绝 H_0''.

两个正态分布总体方差比的检验 设定假设

(1) $H_0 : \dfrac{\sigma_1^2}{\sigma_2^2} \geqslant b \leftrightarrow H_1 : \dfrac{\sigma_1^2}{\sigma_2^2} < b$

(2) $H_0' : \dfrac{\sigma_1^2}{\sigma_2^2} \leqslant b \leftrightarrow H_1' : \dfrac{\sigma_1^2}{\sigma_2^2} > b$

(3) $H_0'' : \dfrac{\sigma_1^2}{\sigma_2^2} = b \leftrightarrow H_1'' : \dfrac{\sigma_1^2}{\sigma_2^2} \neq b$

考虑统计量 $F = \dfrac{S_1^2}{bS_2^2} \sim F_{m-1,n-1}$，水平为 α 的检验为

1. φ：当 $F < F_{m-1,n-1}(1-\alpha)$ 时拒绝 H_0，否则不能拒绝 H_0；
2. φ'：当 $F > F_{m-1,n-1}(\alpha)$ 时拒绝 H_0'，否则不能拒绝 H_0'；
3. φ''：当 $F < F_{m-1,n-1}\left(1-\dfrac{\alpha}{2}\right)$ 时拒绝 H_0''，否则不能拒绝 H_0''.

比例 p 的检验 设 $\boldsymbol{X}$ 是 $0-1$ 分布总体 $B(1,p)$ 的一个样本. 关于 p 设定假设

(1) $H_0 : p \leqslant p_0 \leftrightarrow H_1 : p > p_0$

(2) $H_0' : p \geqslant p_0 \leftrightarrow H_1' : p < p_0$

(3) $H_0'' : p = p_0 \leftrightarrow H_1'' : p \neq p_0$

考虑统计量 $X = \sum\limits_{i=1}^{n} X_i \sim B(n,p)$，由分布函数 $F_p(C) = \sum\limits_{i=1}^{C} C_n^i p^i (1-p)^{n-i}$. 常见的对应的三个检验方法为

1. ψ：当 $X > C$ 时拒绝 H_0，否则不能拒绝 H_0. 其中 C 由下式决定
$$F_{p_0}(C)1-\alpha$$
2. ψ'：当 $X < C'$ 时拒绝 H_0'，否则不能拒绝 H_0'. 其中 C' 由下式决定
$$F_{p_0}(C'-1) = \alpha$$
3. ψ''：当 $C_1 \leqslant X \leqslant C_2$ 时拒绝 H_0''，否则不能拒绝 H_0''. 其中 C_1, C_2 由下式决定
$$F_{p_0}(C_1-1) = \frac{\alpha}{2}, \quad F_{p_0}(C_2) = 1 - \frac{\alpha}{2}$$

以检验 ψ 为例，实际上 C 作为整数不一定能取到，即 C 满足

$$F_{p_0}(C) < 1-\alpha < F_{p_0}(C+1)$$

一种常见的随机化检验是 ψ_0：当 $X \leqslant C$ 时不拒绝 H_0，当 $X > C+1$ 时拒绝 H_0，当 $X = C+1$ 时，从 $[0,1]$ 中任取一个随机数 u，若下式成立则拒绝 H_0 否则不拒绝：

$$u > \frac{1-\alpha-F_{p_0}(C)}{F_{p_0}(C+1)-F_{p_0}(C)}$$

似然比的检验 考虑假设

$$H_0 : \theta \in \Theta_0 \leftrightarrow H_1 : \theta \in \Theta_1 = \Theta / \Theta_0$$

设样本 $\boldsymbol{X}$ 有联合概率质量函数 $f(\boldsymbol{x};\theta)$，称下面统计量为上述假设的似然比

$$LR(\boldsymbol{x}) = \frac{\sup_{\theta\in\Theta} f(\boldsymbol{x};\theta)}{\sup_{\theta\in\Theta_0} f(\boldsymbol{x};\theta)}$$

而一个似然比的检验可以写为：ϕ：当 $LR(\boldsymbol{x}) > c$ 时拒绝原假设 H_0，否则不能拒绝 H_0. 其中常数 c 由检验水平决定（此时可以通过 $LR(\boldsymbol{x})$ 的分布确定）

似然比的极限分布 设 $\dim\Theta - \dim\Theta_0 = t$，则在原假设 H_0 成立之下，当样本量 $n \to \infty$ 时有

$$P(2\ln LR(\boldsymbol{X}) \leqslant x) = F_{\chi_t^2}(x), \quad \forall x \in \mathbb{R}$$

p 值 抽象的定义 p 值为 $P($ 得到和当前样本下检验统计量 T 之值一样或更计算值 $|$ 原假设下 $)$，对应的检验可以写为 ϕ：当 p 值 $< \alpha$ 时，拒绝原假设.

例如正态总体均值的检验 $H_0 : \mu = 2 \leftrightarrow H_1 : \mu > 2$，方差 $\sigma^2 = 1$ 已知，当前样本 $\boldsymbol{X}$ 的均值为 3.5. 和上述一样考虑检验统计量 $T = \dfrac{\sqrt{n}(\bar{X}-2)}{1}$，当前样本下检验统计量为 $t_{\text{obs}=3.5}$，故对应的 p 值写为 $P(T \geqslant t_{\text{obs}} | H_0) = 1 - \Phi(t_{\text{obs}})$，其中 H_0 即代表 $\mu = 2$ 我们已代入检验统计量中.

非参数假设检验

理论分布完全已知且有限个取值的拟合优度检验 总体 $\boldsymbol{X}$ 的值域为 $\{a_1, a_2, \ldots, a_k\}$，抽取一个样本量为 n 的简单样本，其中有 n_i 次取 a_i，则

$$Z = \sum_{i=1}^{k} \frac{n_i^2}{np_i} - n$$

在原假设 $H_0 : P(X = a_i) = p_i, \forall i \leftrightarrow H_1 : \exists j, P(X = a_j) \neq p_j$ 成立时，当 $n \to \infty$，Z 的分布趋于自由度为 $k-1$ 的 χ^2 分布. 对应的检验写为：

φ：当 $Z > \chi_{k-1}^2(\alpha)$ 时拒绝 H_0，否则不能拒绝 H_0.

同时定义数据对理论分布的“拟合优度”

$$p(Z_0) = P(Z \geqslant Z_0) = 1 - F_{\chi_{k-1}^2}(Z_0)$$

理论分布类型已知但含有有限个未知参数 总体 $\boldsymbol{X}$ 的值域为 $\{a_1, a_2, \ldots, a_k\}$，但含有 r 个未知参数 $\theta_1, \theta_2, \ldots, \theta_n$ $(r < k-1)$. 抽取一个样本量为 n 的简单样本，其中有 n_i 次取 a_i，则

$$Z = \sum_{i=1}^{k} \frac{n_i^2}{np_i} - n$$

在原假设 $H_0' : P(X = a_i) = p_i(\theta_1, \theta_2, \ldots, \theta_n), \forall i$ 成立时，当 $n \to \infty$，Z 的分布趋于自由度为 $k-r-1$ 的 χ^2 分布. 对应的检验写为：

ϕ：当 $Z > \chi_{k-r-1}^2(\alpha)$ 时拒绝 H_0，否则不能拒绝 H_0.

列联表检验 检验属性 A, B 独立的假设 H_0，我们设 A, B 分别处于水平 i, j 的样本数量为 $n_{i,j}$，且 $n_{i\cdot} = \sum\limits_{k=1}^{b} n_{ik}$, $n_{\cdot j} = \sum\limits_{k=1}^{a} n_{kj}$，则可以写出检验统计量

$$Z = \sum_{i=1}^{a}\sum_{j=1}^{b} \frac{(nn_{ij} - n_{i\cdot}n_{\cdot j})^2}{nn_{i\cdot}n_{\cdot j}} \sim \chi_{(a-1)(b-1)}^2$$

对应的检验写为

ψ：当 $Z > \chi_{(a-1)(b-1)}^2(\alpha)$ 时拒绝 H_0，否则不能拒绝 H_0.

随机变量分布及其数字特征

离散随机变量

分布名称	概率质量函数 (pmf)	期望	方差
$0-1$ 分布	$P(X=1)=p$	p	
二项分布 $X\sim B(n,p)$	$P(X_r=k)=C_n^k p^k(1-p)^k=b(n,p,k)$	np	$np(1-p)$
负二项分布（Pascal 分布）$X\sim NB(n,p)$	$P(X=k)=C_{k-1}^{r-1}p^r(1-p)^{k-r}=nb(r,p,k)$	$\dfrac{n}{p}$	
Poisson 分布 $X\sim P(\lambda)$	$P(X=k)=\mathrm{e}^{-\lambda}\dfrac{\lambda^k}{k!}$	λ	λ

连续随机变量

分布名称	概率密度函数 (pdf)	期望	方差
均匀分布 $X\sim U(a,b)$	$f(x)=\dfrac{1}{b-a}I_{(a,b)}(x)$	$\dfrac{a+b}{2}$	$\dfrac{b-a}{12}$
指数分布 $X\sim Exp(\lambda)$	$f(x)=\lambda\mathrm{e}^{-\lambda}I_{(0,\infty)}(x)$	λ^{-1}	λ^{-2}
正态分布 $X\sim N(\mu,\sigma^2)$	$f(x)=\dfrac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\dfrac{(x-\mu)^2}{2\sigma^2}\right\},\ x\in\mathbb{R}$	μ	σ^2
Cauchy 分布	$f(x)=\dfrac{1}{\pi(1+x^2)}$	不存在	

多维（二元）随机变量

二元分布名称	联合概率密度函数 (pdf)	期望	方差	相关系数
均匀分布	$f(x,y)=\dfrac{1}{\lvert G\rvert}I_G(x,y)$，$G$ 为面积为 $\lvert G\rvert\neq 0$ 的有界区域	/	/	/
二元正态分布 $X\sim N(a,b,\sigma_1^2,\sigma_2^2,\rho)$	$f(x,y)=\dfrac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\left\{-\dfrac{1}{2(1-\rho^2)}\left[\dfrac{(x-a)^2}{\sigma_1^2}-2\rho\dfrac{(x-a)(y-b)}{\sigma_1\sigma_2}+\dfrac{(y-b)^2}{\sigma_2^2}\right]\right\}$	$\mathrm{E}X=a,\ \mathrm{E}Y=b$	$\mathrm{Var}X=\sigma_1^2,\ \mathrm{Var}Y=\sigma_2^2$	ρ
n 元正态分布	$f(\boldsymbol{x})=\dfrac{1}{\sqrt{(2\pi)^n\lvert\boldsymbol{A}\rvert}}\exp\left\{-\dfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{A}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\},\quad A_{ij}=\rho_{ij}\sigma_i\sigma_j$	$\mathrm{E}X_i=\mu_i$	$\mathrm{Var}X_i=A_{ii}$	$\mathrm{Cov}(X_i,X_j)=\dfrac{A_{ij}}{\sigma_i\sigma_j}$

课本涉及的典型问题

第一章：事件及其概率

1. Bertrand 悖论
2. 敏感性问题调查
3. 波利亚罐子模型
4. 核酸检测 (Bayes)
5. 小概率事件
6. 两两独立而不相互独立
7. 递推法
8. 配对问题
9. 贝叶斯公式与垃圾邮件识别
10. 三门问题

第二章：随机变量及其分布

1. 数字通信及可靠性
2. 标记重捕模型
3. Banach 火柴问题
4. 负二项分布的 Poisson 近似
5. Weibull 分布

第三章：所谓随机变量及其分布

1. 会面问题
2. 独立随机变量的和（卷积）
3. 指数分布随机变量的和与差
4. 正态分布随机变量的和
5. 独立随机变量商的商
6. Cauchy 分布
7. 最大值和最小值的分布
8. 系统可靠性研究
9. Simpson 悖论

第四章：随机变量的数字特征和极限定理

1. 巴格达窃贼问题
2. 随机变量标准化
3. 偏度系数和峰度系数
4. 超几何分布的期望
5. 配对问题
6. 游程问题

第五章：统计学基本概念

1. $X\sim U(0,\theta)$ 总体抽取的简单样本中，统计量 $X_{(n)}$ 的抽样分布
2. Γ 分布

第六章：参数点估计

1. 总体标准差的无偏估计
2. 德军坦克问题

第七章：区间估计

1. “足球赛会杀人”巧合

、

第八章：假设检验

1. 符号检验
2. 置信区间和假设检验之间的关系
3. 多重假设检验

⋆ 祝考试顺利！ ⋆