懂你的 gay 捅

xjj

高中概率知识省略

*表示拓展内容,个人认为考察概率不大

概率论部分

一）随机变量与分布函数

1）分布函数

$$def:\ F: R \to [0,1], F(x) = P(X \le x_k)\ , F\text{右连续}$$

2） 离散型随机变量

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

→分布

1）二项分布(事件 A 在 n 次试验中发生的次数)

$$X \sim B(n, p)$$

$$P_n(k) = P(X = k) = C_n^k p^k (1-p)^{n-k}, k = 0,1,\dots,n$$

2）*负二项(pascal)分布(事件 A 第 r 次发生时的试验次数)

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

3）超几何分布(不放回抽样 n 个,总数 N,对应样本 M)

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

4）泊松(Poisson)分布(稀有事件在大量重复试验中出现的次数)

$$X \sim \pi(\lambda) \mathrm{or}\ P(\lambda)$$

$$P(X = k) = \frac{e^{-\lambda} \lambda^k}{k!}\ , k = 0,1,\dots$$ (可查表)

$\lim\limits_{N \to +\infty}$ 超几何分布=二项分布

$\lim\limits_{n \to +\infty}$ 二项分布=泊松分布 (泊松定理)

3）连续型随机变量

→概率密度

$$def:\ \int_{-\infty}^{+\infty} f(x)dx = 1 \quad f(x) \ge 0$$

→推论(求解)

$$cor: f(x) = F'(x)$$

→分布

1） 均匀分布

$$X\sim U(a,b)\quad f(x)=\frac{1}{b-a}\quad x\in(a,b)$$

2） 指数分布(通常作为某种“寿命”的近似|无记忆性)

$$X\sim E(\lambda),f(x)=\lambda e^{-\lambda x},F(x)=1-e^{-\lambda x},x\geq 0$$

3） 正态(Gauss)分布

$$X\sim N(\mu,\ \sigma^2)\ \ f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$F(x)=\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{x}e^{-\frac{(t-\mu)^2}{2\sigma^2}}dt$$

在$x=\mu\pm\sigma$处，$f(x)$有拐点

→退化$N(1,0)$,$\varphi(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$

$$\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{t^2}{2}}dt$$ (可查表)

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

→3σ法则

→二项分布的近似计算

$$P(a<X<b)\approx\int_{x_1}^{x_2}\varphi(x)dx,$$

$$x_1=\frac{a-np}{\sqrt{np(1-p)}},x_2=\frac{b-np}{\sqrt{np(1-p)}}$$

4） *Γ分布

$$X\sim\Gamma(\alpha,\beta)$$

$$\Gamma(\alpha)=\int_0^{+\infty}x^{\alpha-1}e^{-x}dx$$

$$f(x)=\frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}(x>0)$$

$$\Gamma(\alpha+1)=\alpha\Gamma(\alpha)\quad\Gamma(0.5)=\sqrt{\pi}$$

$$n\in N\Rightarrow\Gamma(n+1)=n!$$

$$\frac{\Gamma\left(\frac{k_1}{2}\right)\Gamma\left(\frac{k_2}{2}\right)}{\Gamma\left(\frac{k_1+k_2}{2}\right)}=\int_0^1 x^{\frac{k_1}{2}-1}(1-x)^{\frac{k_2}{2}-1}dx$$

→退化 1 $\Gamma(1,\beta)=\mathrm{E}(\beta)$(指数分布)

→退化 2 $\Gamma(n/2,1/2)$ (χ^2分布,记作$\chi^2(n)$)

4)随机变量的函数

$$Y=g(X)$$

→离散型随机变量

$$P(Y=y_i)=\sum_{k:g(x_k)=y_i}p_k$$

→连续型随机变量

二）多维随机变量及其概率分布

1)多维随机变量

$$def:\forall\omega\in\Omega\xrightarrow{一定法则}\exists\big(X_1(\omega),X_2(\omega),\dots,X_n(\omega)\big)\in R^n,$$

$$X=(X_1,X_2,\dots,X_n)$$

2)联合分布函数/边缘分布函数

$$F(x_1,x_2,\dots,\ x_n)=P(X_1\le x_1,X_2\le x_2,\dots,X_n\le x_n)$$

将上联合分布函数保留 k 个 X 的分量,其他替换为+∞并省略,即为边缘分布函数

对$f(x_1,x_2,\dots,\ x_n)$关于其他分量的积分即为边缘概率分布

3)二维随机变量

离散型 $P(X=x_i,Y=y_j)=p_{ij}\ge 0\ ,i,j=1,2,\dots,\ \sum p_{ij}\ =1$

连续型 $\iint f(x,y)dxdy=1\ ,\frac{\partial^2 F}{\partial x\,\partial y}=f(x,y)\ge 0$

4)二维均匀分布

$f(x,y)=\frac{1}{A},(x,y)\in G$ ，记作$U(G)$.

5)二维正态分布

$$(X,Y)\sim N(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$$

$$f(x,y)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-p^2}}e^{-\frac{1}{2(1-p^2)}\left[\frac{(x-\mu_1)^2}{\sigma_1}-\frac{2\rho(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2}+\frac{(y-\mu_2)^2}{\sigma_2}\right]}$$

二维正态分布的边缘分布仍为正态分布。

6)n 维正态分布

$$f(x)=\frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}}e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)},$$

$$\mu=\binom{\mu_1}{\mu_2},\Sigma=\begin{bmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{bmatrix}$$

7)随机变量的独立性

若对$\forall x,y$都有:$P(X\le x,Y\le y)=P(X\le x)P(Y\le y)$

则称随机变量X与Y相互独立

→结论 1

cor:若存在非负可积函数$r(x),g(x)$,使得

$f(x,y)=r(x)g(y)(a.e.)$ 则X与Y相互独立

→结论 2

cor: X与Y相互独立的充要条件为$F(x,y)=R(x)G(y)$

三）随机变量的条件分布

1)离散型随机变量

$\frac{P(X=x_i,Y=y_i)}{P(Y=y_i)}=P(X=x_i|Y=y_i)$ 为在$Y=y_i$的条件下X的条件概率分布

$P(A)=\sum P(B_i)P(A|B_i)$ 全概率公式

$P(B_k|A)=\frac{P(B_k)P(A|B_k)}{\sum P(B_i)P(A|B_i)}$Bayes 公式

2)一般条件下条件分布

$$F_{X|Y}(x|y)=\lim_{\varepsilon\to 0^+}P\{X\le x|y-\varepsilon<Y\le y+\varepsilon\}$$

为在$Y=y_i$的条件下X的条件分布函数

$f_{X|Y}(x|y)=\frac{f(x,y)}{f_Y(y)}$为在$Y=y_i$的条件下$X$的条件分布密度

类似可得全概率公式与 Bayes 公式

3)*求$f_{ZU}(z,u)$

设$\begin{cases}z=g(x,y)\\u=r(x,y)\end{cases}$存在唯一的反函数:$\begin{cases}x=h(z,u)\\y=s(z,u)\end{cases}$

且h,s有连续偏导数,并记雅可比行列式$J(z,u)=\begin{vmatrix}\frac{\partial h}{\partial z} & \frac{\partial h}{\partial u}\\ \frac{\partial s}{\partial z} & \frac{\partial s}{\partial u}\end{vmatrix}$

则$f_{ZU}(z,u)=f_{XY}\big(h(z,u),s(z,u)\big)|J|$

4)*和的分布

$$Z=X+Y\quad f_Z(z)=\int_{-\infty}^{+\infty}f_X(x)f_Y(z-x)dx$$

称之为$f_X(x)$与$f_Y(y)$的卷积。

涉及正态随机变量结论:

→$(X,Y)\sim N(\mu_1,\sigma_1^2;\mu_1,\sigma_1^2;\rho)$

$$\Rightarrow X+Y\sim N(\mu_1+\mu_2,\sigma_1^2+2\rho\sigma_1\sigma_2+\sigma_2^2)$$

→$X_1,X_2,\dots X_n$相互独立$\Rightarrow\sum X\sim N(\sum\mu_1,\sum\sigma_i^2)$

5)*商的分布

$$Z=X/Y\quad f_{X/Y}(z)=\int_{-\infty}^{+\infty}f(zy,y)|y|dy$$

6)*平方和的分布

$$Z=X^2+Y^2\quad f_Z(z)=\begin{cases}0,\ z<0,\\ \frac{1}{2}\int_0^{2\pi}f\big(\sqrt{z}cos\theta,\sqrt{z}sin\theta\big)d\theta,z\geq 0\end{cases}$$

7)极值函数的分布

$$M=\max\{X,Y\}\quad F_M(u)=F_X(u)F_Y(u)$$

$$N=\min\{X,Y\}\ F_N(u)=1-[1-F_X(u)][1-F_Y(u)]$$

四）随机变量的数字特征

1)数字特征的定义与存在判据

→期望与方差

$def:Var(X)=D(X)=E[X-E(X)]^2,\sigma=\sqrt{D(X)}$为均方差/标准差

$$D(X)=E(X^2)-(E(X))^2$$

存在判据:自身绝对收敛

→*偏度系数与峰度系数

$$\alpha=\frac{E\big(X-E(X)\big)^3}{\big(D(X)\big)^{\frac{3}{2}}}\ ,\gamma=\frac{E\big(X-E(X)\big)^4}{\big(D(X)\big)^2}$$

存在判据:分子

α:刻画随机变量取值关于其数学期望的对称程度

γ:刻画随机变量在期望和方差确定时其概率分布的峰态。如对连续型随机变量，刻画其密度函数曲线的陡峭状态。

→*变异系数

$$\nu = \frac{\sqrt{D(X)}}{E(X)}$$

存在判据:非负随机变量&$E(X) > 0$&$D(X)$存在

ν:反映随机变量的离散程度且无量纲

→分位数

若$P(X \le a) \ge p \ge P(X < a)$,则称$a$为$X$的$p$分位数,记为$x_p = a$

$x_{0.5} = med(X)$

注意到x_p不唯一,若定义$\begin{matrix} x_p = \sup\{a: P(X \le a) < p\} \\ or = \inf\{a: P(X \le a) \ge p\} \end{matrix}$,则唯一

def:上下侧分位数

→标准化随机变量

$$X^* = \frac{X - E(X)}{\sqrt{D(X)}}$$

→协方差与相关系数

协方差$cov(X,Y) = E\Big(\big(X - E(X)\big)\big(Y - E(Y)\big)\Big) = E(XY) - E(X)E(Y)$

$$\begin{cases} \rho_{XY} = 1 \Leftrightarrow Y = aX + b \\ \rho_{XY} = 0 \Leftrightarrow X, Y\text{不相关} \end{cases}$$

X, Y相互独立 $\Rightarrow$ X, Y 不相关 , X, Y相互独立 $\xLeftarrow{\text{二维正态分布}}$ X, Y 不相关

→协方差矩阵

称$\Sigma(x_1, x_2, \dots, x_n) = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{bmatrix}$为$(x_1, x_2, \dots, x_n)$

的协方差矩阵,其中$\sigma_{ij} = cov(X_i, X_j)$, 该矩阵对称非负定

→*矩

$E(X^k)$: X的k阶原点矩

$E(|X|^k)$: X的k阶绝对原点矩

$E\left(\big(X - E(X)\big)^k\right)$: X的k阶中心矩

$E(X^k Y^l)$: X, Y的$k + l$阶混合原点矩

$E\left(\big(X - E(X)\big)^k \big(Y - E(Y)\big)^l\right)$: X, Y的$k + l$阶混合中心矩

2）离散型$E(X)=\sum_{k=1}^{+\infty} x_k p_k$

→$X\sim B(n,p) E(X)=np \ \ D(X)=np(1-p)$

→$X\sim P(\lambda) \quad\quad E(X)=D(X)=\lambda$

→$X\sim$超几何分布$(N,M,n) \ \ E(X)=\frac{nM}{N} \ \ D(X)=\frac{nM}{N}\left(1-\frac{M}{N}\right)\frac{N-n}{N-1}$

→$X\sim$几何分布$(p) \ \ E(X)=\frac{1}{p} \ \ D(X)=\frac{1-p}{p^2}$

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

→$X\sim U(a,b) \ \ E(X)=\frac{a+b}{2} \ \ D(X)=\frac{(b-a)^2}{12}$

→$X\sim E(\lambda) \ \ E(X)=\frac{1}{\lambda} \ \ D(X)=\frac{1}{\lambda^2}$

→$X\sim N(\mu,\ \sigma^2) \ E(X)=\mu \ \ D(X)=\sigma^2$

→$X\sim \Gamma(\alpha,\beta) \ \ E(X)=\frac{\alpha}{\beta} \ \ D(X)=\frac{\alpha}{\beta^2}$

4)$Y=g(X)$

$$E(Y)=\sum_{k=1}^{+\infty} g(x_k)p_k$$

$$E(Y)=\int_{-\infty}^{+\infty} g(x)f(x)dx$$

5)$Z=g(x,y)$的数学期望(可推广至 n 元)

$$E(Z)=\sum_{i,j=1}^{+\infty} g\left(x_i y_j\right)p_{ij}$$

$$E(Z)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} g(x,y)f(x,y)dx\,dy$$

6)一些性质

$|E(XY)|^2\le E(X^2)E(Y^2)$ Schwarz 不等式

$D(X\pm Y)=D(X)+D(Y)\pm 2E\left(\left(X-E(X)\right)\left(Y-E(Y)\right)\right)$

X,Y相互独立$\Rightarrow D(X\pm Y)=D(X)+D(Y)$

$D(X)\le E(X-C)^2$

五）概率极限理论

1)随机变量序列的收敛性

→概率为 1 的收敛

$P\{\lim_{n\to\infty} X_n = X\} = 1$,记作$X_n \xrightarrow{a.e} X$又称几乎处处(几乎必然)收敛于$X$

→依概率收敛

$\forall\varepsilon > 0, \lim_{n\to\infty} P\{|X_n - X| < \varepsilon\} = 1$,记作$X_n \xrightarrow{P} X$

→依分布收敛(弱收敛)

对$F(x)$所有连续点x, $\lim_{n\to\infty} F_n(x) = F(x)$,记作$X_n \xrightarrow{d} X$

→定理:

$X_n \xrightarrow{a.e} X \Rightarrow X_n \xrightarrow{P} X \Rightarrow X_n \xrightarrow{d} X$

2)大数定律

若$E(X)$存在,则对$\forall\varepsilon > 0, P(X \geq \varepsilon) \leq \frac{E(X)}{\varepsilon}$

→Markov 不等式

若$E(|X|^k)$存在,则对$\forall\varepsilon > 0, P(|X| \geq \varepsilon) \leq \frac{E(|X|^k)}{\varepsilon^k}$

→Chebyshev 不等式

若$D(X)$存在,则对$\forall\varepsilon > 0, P(|X - E(X)| \geq \varepsilon) \leq \frac{D(X)}{\varepsilon^2}$

→算术平均与大数定律

若$E(X_k)$存在,则记$\overline{X_n} \stackrel{\text{def}}{=} \frac{1}{n}\sum_{k=1}^{n} X_k$

$\overline{X_n} \xrightarrow{P} E(\overline{X_n})$,则称序列$\{X_k\}$服从大数定律

→Bernoulli 大数定律:X_k为独立重复实验(即相互独立的 0-1 分布)(依概率稳定)

→Chebyshev 大数定律:X_k相互独立且具有相同期望和方差

→*相互独立条件可去

→*Khinchin 大数定律

→*强大数定律

→柯尔莫哥洛夫强大数定律(Khinchin 大数定律增强)

→波雷尔 Borel 强大数定律(Bernoulli 大数定理增强)

3)中心极限定理

若$E(X_k)D(X_k)$都存在,且有$\frac{n\overline{X_n} - E(n\overline{X_n})}{\sqrt{D(n\overline{X_n})}} \xrightarrow{d} X \sim N(0,1)$,

则称序列$\{X_k\}$服从中心极限定理

(即随机变量$n\overline{X_n} = \sum_{k=1}^{n} X_k$的标准化变量近似$\sim N(0,1)$)

(即$\sum_{k=1}^{n} X_k \sim N(n\mu, n\sigma^2)$)

→独立同分布中心极限定理

→De Moivre-Laplace 中心极限定理(二项分布,并说明正太分布是其极限分布律)

数理统计部分

→经验分布函数

Y_n的分布函数F_n(对总体X样本值$x_1 \le x_2 \le \cdots \le x_n, Y_n$等可能地取到这$n$个值的每一个)

→格列汶科定理

$$P\left\{\lim_{n\to\infty} \overset{sup}{-\infty < x < +\infty} |F_n(x) - P(X \le n)| = 0\right\} = 1$$

→常用统计量:

样本均值$\bar{X}$/k阶原点矩$A_k\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right)$/$k$阶中心距$B_k\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right)$/样本方差

$S^2\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\right)$/样本标准差$S$/$B_2 = S_n^2$

→顺序统计量与极差$D_n = X_{(n)} - X_{(1)}$

一）　来自正态总体常用统计量及其分布

1）$\chi^2(\mathrm{n})$分布(卡方分布)

设$X_1, X_2, \ldots, X_n$为标准正态总体$N\ (0,1)$的样本，则称如下统计量为χ^2统计量:

$$\sum_{i=1}^{n} X_i^2 \sim \chi^2(\mathrm{n})$$

$n = 2$时为参数为 1/2 的指数分布,$n = 1$ 时,$f(x) = \begin{cases} \frac{1}{\sqrt{2\pi}} x^{-\frac{1}{2}} e^{-\frac{x}{2}} & x>0 \\ 0 & x\le 0 \end{cases}$

$\rightarrow E(\chi^2(\mathrm{n})) = \mathrm{n}, \mathrm{D}(\chi^2(\mathrm{n})) = 2\mathrm{n}$

→可加性$X_1 + X_2 \sim \chi^2(\mathrm{n_1} + \mathrm{n_2})$

$\rightarrow \lim_{n\to+\infty} \chi^2(\mathrm{n}) =$ 正态分布

2）$t(n)$分布(Student 分布)

设$X \sim N(0,1), Y \sim \chi^2(\mathrm{n}), X, Y$相互独立,则称如下统计量为$t$统计量:

$$T = \frac{X}{\sqrt{\frac{Y}{n}}} \sim t(n)$$

$\rightarrow \lim_{n\to+\infty} f_n(t) \sim \mathrm{N}(0,1)$

$\rightarrow t_{1-\alpha}(n) + t_\alpha(n) = 0$

3）$F(n,m)$分布(n,m分别为第一、二自由度)

设$X \sim \chi^2(\mathrm{n}), Y \sim \chi^2(\mathrm{m}), X, Y$相互独立,则称如下统计量为$F$统计量:

$$F = \frac{\frac{X}{n}}{\frac{Y}{m}} \sim F(n,m)$$

$\rightarrow \frac{1}{F} \sim F(m,n)$

$\rightarrow F_{1-\alpha}(n,m) F_\alpha(m,n) = 1$

$\rightarrow t_{\frac{\alpha}{2}}^2(n) = F_\alpha(1,n)$

4) 正态总体 7 大结论

$\rightarrow \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

$\rightarrow \frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{\sigma}\right)^2 \sim \chi^2(n-1)$

$\rightarrow \frac{(n-1)S^2}{\sigma^2}$与$\bar{X}$相互独立

$\rightarrow \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} / \frac{S}{\sigma} = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$

$\rightarrow \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n-1, m-1)$

$\rightarrow$对于$\sigma_1 = \sigma_2, \frac{S_1^2}{S_2^2} \sim F(n-1, m-1)$

$\rightarrow \frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{1}{n}+\frac{1}{m}}\sqrt{\frac{(n-1)S_1^2+(M-1)S_2^2}{n+m-2}}} \sim t(n+m-2)$

二) 参数估计

1) 点估计的思想方法

设总体X的分布函数的形式已知，但含有k个未知参数:$\theta_1, \theta_2, \ldots, \theta_k$,基于总体的一个样本$X_1, X_2, \ldots, X_n$构造$k$个统计量:

$\hat{\theta}_1(X_1, X_2, \ldots, X_n)\hat{\theta}_2(X_1, X_2, \ldots, X_n)\cdots\hat{\theta}_k(X_1, X_2, \ldots, X_n)$即矩估计量

代入$(x_1, x_2, \ldots, x_n)$得到k个估计值

$\rightarrow$频率替换法 $\hat{p}_A = \frac{n_A}{n}$ 依据:伯努利大数定律$\frac{n_A}{n} \xrightarrow{P} p_A$

$\rightarrow$令$p_A = p_A(\theta)$,利用$p_A(\hat{\theta}) = \frac{n_A}{n}$

2) 矩估计

$\hat{\mu}_k = A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k, \hat{G} = g(\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_n)$

$\rightarrow$依据:推广版辛钦定理:$A_r \xrightarrow{P} \mu_r$

$\rightarrow$K 个矩估计量的求解

即求解方程组$\mu_r(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_n) = \frac{1}{n}\sum_{i=1}^{n} X_i^k, r = 1,2, \ldots, k$

$e.g. \hat{\mu} = \bar{X}, \hat{\sigma}^2 = S_n^2$

3) 极(最)大似然估计法

$\rightarrow$似然函数:

设总体X的密度函数(或概率分布)为$f(x_i, \theta), \theta \in \Theta, \Theta$为$\theta$可能取值的集合(或范围),则对于来自总体 X 的简单随机样本$(X_1, X_2, \ldots, X_n)$,其联合密度函数(或联合概率分布)为:

$L(x_1, x_2, \ldots, x_n, \theta) = \prod_{i=1}^{n} f(x_i, \theta)$

当样本$(X_1, X_2, \ldots, Xn)$观测值$(x1, x2, \ldots, xn)$给定时,退变为θ的函数,记为:$L(\theta)$.称$L(\theta)$为样本的似然函数.

$\rightarrow$极大似然估计

选择适当的$\theta=\hat{\theta}$,使$L(\theta)$取到最大值,即:

$$L\left(x_1,x_2,\dots,x_n,\hat{\theta}\right)=\max_{\theta\epsilon\Theta}\{L(x_1,x_2,\dots,x_n,\theta)\}$$

得到$\hat{\theta}=g(x_1,x_2,\dots,x_n)$

→此处θ可扩充为$(\theta_1,\theta_2,\dots,\theta_k)$

→似然方程:$\frac{\partial L(\theta)}{\partial\theta}|_{\theta=\hat{\theta}}=0,\frac{\partial^2L(\theta)}{\partial\theta^2}|_{\theta=\hat{\theta}}<0,or$对数似然方程$\frac{\partial lnL(\theta)}{\partial\theta}|_{\theta=\hat{\theta}}=0$

→依据:一次试验(取一个样本)中,所出现的事件有较(最)大的概率.

→极大似然估计的不变性:设$\hat{\theta},u(\theta)$,单值函数$\theta(u)$,则有 $\hat{u}=u\left(\hat{\theta}\right)$

三） 估计量的评价

1)无偏性

对 $\hat{\theta}(X_1,X_2,\dots,X_n)$,若$E\left(\hat{\theta}\right)=\theta$,则称$\hat{\theta}$是$\theta$的无偏估计量.

→样本原点矩是总体原点矩的无偏估计

2)有效性

若$D\left(\hat{\theta}_1\right)<D\left(\hat{\theta}_2\right)$,则称 $\hat{\theta}_1$比$\hat{\theta}_2$更有效

→算术均值比加权均值更有效

3)一致性

若 $\hat{\theta}\xrightarrow{P}\theta$,则称 $\hat{\theta}$是总体参数θ的一致/相合估计量

→样本 k 阶矩是总体 k 阶矩的一致性估计量(由大数定律可知)

→设 $\hat{\theta}$是θ的无偏估计量,且$\lim\limits_{n\to\infty}D\left(\hat{\theta}\right)=0$,则$\hat{\theta}$是$\theta$的一致估计量(chebyshev)

四） 区间估计

记 $u_\alpha,t_\alpha(n),\chi_\alpha^2(n),F_\alpha(n,m)$分别为标准正态分布,$t(n)$分布,$\chi^2(n)$分布,$F(n,m)$分布的α下侧分位数

1）基本步骤

A．寻找样本的函数:

$$g(X_1,X_2,\dots,X_n,\theta)$$

称为枢轴量

含有待估参数、不含其它未知参数，分布已知，且不依赖于待估参数

B．对于给定的置信度$1-\alpha$，确定出(用以构造随机事件的)常数a和b,使得:

$$P(a<g(X_1,X_2,\dots,X_n,\theta)<b)=1-\alpha$$

C．由$a<g(X_1,X_2,\dots,X_n,\theta)<b$解出$\theta$的置信上下限:

$$\hat{\theta}_1(X_1,X_2,\dots,X_n)\ \hat{\theta}_2(X_1,X_2,\dots,X_n)$$

D．得到置信区间:

$$\left(\hat{\theta}_1,\hat{\theta}_2\right)$$

2)一个正态总体 $X\sim N\ (\mu\ ,\sigma^2\ )$的情形

→方差σ^2已知,μ的置信区间

$$\left(\bar{X}-u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}},\bar{X}+u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)\quad \text{枢轴量}\frac{(\bar{X}-\mu)}{\sigma/\sqrt{n}}\sim N(0,1)$$

→方差σ^2未知,μ的置信区间

$$\left(\bar{X}-t_{1-\frac{\alpha}{2}}(n-1)\frac{S}{\sqrt{n}},\bar{X}+t_{1-\frac{\alpha}{2}}(n-1)\frac{S}{\sqrt{n}}\right)\ \text{枢轴量}\frac{\bar{X}-\mu}{S/\sqrt{n}}\sim t(n-1)$$

→ μ已知,方差σ^2的置信区间

$$\left(\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{1-\frac{\alpha}{2}}(n)},\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{\frac{\alpha}{2}}(n)}\right)\ \text{枢轴量}\sum_{\mathrm{i=1}}^{\mathrm{n}}\left(\frac{X_i-\mu}{\sigma}\right)^2\sim\chi^2(n)$$

→ μ未知,方差σ^2的置信区间

$$\left(\frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)},\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}\right)\ \text{枢轴量}\frac{(n-1)S^2}{\sigma^2}\sim\chi^2(n-1)$$

3)两个正态总体的情形

设$(X_1,X_2,\dots,X_n)$和$(Y_1,Y_2,\dots,Y_m)$分别为来自总体$N(\mu_1,\sigma_1^2)$和总体$N(\mu_2,\sigma_2^2)$的相互独立的样本,$\bar{X},S_1^2;\bar{Y},S_2^2$分别为相应的样本均值与样本方差

→ $\sigma_1^2\sigma_2^2$已知,$\mu_1-\mu_2$的置信区间

$$\left((\bar{X}-\bar{Y})\pm u_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}}\right)\ \text{枢轴量}\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}}}\sim N(0,1)$$

→ $\sigma_1^2\sigma_2^2$未知(但$\sigma_1^2=\sigma_2^2=\sigma^2$),$\mu_1-\mu_2$的置信区间

$$\left((\bar{X}-\bar{Y})\pm t_{1-\frac{\alpha}{2}}(n+m-2)\sqrt{\frac{1}{n}+\frac{1}{m}}\sqrt{\frac{(n-1)S_1^2+(m-1)S_2^2}{n+m-2}}\right)$$

$$\text{枢轴量}\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{1}{n}+\frac{1}{m}}S_w}\sim t(n+m-2)$$

$$\text{其中}\mathrm{S_w}=\sqrt{\frac{(n-1)S_1^2+(m-1)S_2^2}{n+m-2}}$$

→ $\sigma_1^2\sigma_2^2$未知,$n,m>50,\mu_1-\mu_2$的置信区间

$$\left((\bar{X}-\bar{Y})\pm u_{1-\frac{\alpha}{2}}\sqrt{\frac{S_1^2}{n}+\frac{S_2^2}{m}}\right)$$

$$\text{枢轴量}\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{S_1^2}{n}+\frac{S_2^2}{m}}}\frac{S_1^2}{n}+\frac{S_2^2}{m}\underset{\sim}{\approx}\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}N(0,1)$$

→ $\sigma_1^2\sigma_2^2$未知,$n=m,\mu_1-\mu_2$的置信区间

$$\left((\bar{X}-\bar{Y})\pm t_{1-\frac{\alpha}{2}}(n-1)\frac{S_Z}{\sqrt{n}}\right),\text{其中 }\mathrm{S_Z^2}=\frac{\sum_{i=1}^{n}\left[\left(X_i-Y_j\right)-(\bar{X}-\bar{Y})\right]^2}{n-1}$$

$$对Z_i = X_i - Y_i \sim N(\mu_1-\mu_2, \sigma_1^2+\sigma_2^2), 枢轴量\frac{\bar{Z}-\mu_Z}{S_Z/\sqrt{n}}$$

$\rightarrow \frac{\sigma_1^2}{\sigma_2^2}$的置信区间($\mu_1, \mu_2$未知)

$$\left(\frac{S_1^2}{S_2^2}\frac{1}{F_{1-\frac{\alpha}{2}}(n-1,m-1)}, \frac{S_1^2}{S_2^2}\frac{1}{F_{\frac{\alpha}{2}}(n-1,m-1)}\right) \quad 枢轴量F=\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n-1,m-1)$$

$\rightarrow \frac{\sigma_1^2}{\sigma_2^2}$的置信区间($\mu_1, \mu_2$已知)

$$\left[\frac{\frac{m}{n}\frac{\sum_{i=1}^{n}(X_i-\mu_1)^2}{\sum_{j=1}^{m}(Y_j-\mu_2)^2}}{F_{1-\frac{\alpha}{2}}(n,m)}, \frac{\frac{m}{n}\frac{\sum_{i=1}^{n}(X_i-\mu_1)^2}{\sum_{j=1}^{m}(Y_j-\mu_2)^2}}{F_{\frac{\alpha}{2}}(n,m)}\right] \quad 枢轴量\frac{\frac{m}{n}\frac{\sum_{i=1}^{n}(X_i-\mu_1)^2}{\sum_{j=1}^{m}(Y_j-\mu_2)^2}}{\frac{\sigma_1^2}{\sigma_2^2}} \sim F(n,m)$$

4)非正态总体参数的置信区间

利用中心极限定理构建近似分布

5)单侧区间估计

单侧置信区间:$(\underline{\theta}, +\infty)(-\infty, \bar{\theta})$

6)$Bayes$估计

$Bayes$估计观点:待估参数θ是一个随机变量.其估计值看作是该随机变量的实现

利用θ的先验信息对其分布加以表示,即为θ的**先验分布**,记为$\pi(\theta)$.

综合θ的先验信息和样本X带来的信息,得到关于θ的进一步的信息可记为$\pi(\theta|X)$.即为θ的**后验分布**

二者的联系由$Bayes$公式给出;

$$\pi(\theta|X) = \frac{\pi(\theta)\pi(X|\theta)}{\pi(X)}$$

贝叶斯统计的重要意义在于,在统计推断中以概率分布(即先验分布)的形式考虑了关于研究对象的先验信息

五） 假设检验

1)概念

假设阶段:对于总体的某待推断的未知方面,根据解决问题的需要,作出一个假设.

检验阶段:为判断所作的假设是否正确,从总体中抽取样本,并根据样本提供的信息,按一定原则对所作假设加以检验,进而以具体检验情况或实际检验结果作出决定:接受或拒绝所作假设.

称用于检验的统计量V为检验统计量.

2)原假设与备择假设

设θ为总体的一个未知参数,其一切可能值的集合记为Θ.

则关于θ的任一假设可用“$\theta \in \Theta$”来表示,其中Θ^*为Θ的一个真子集.

在统计假设检验中,将在假设阶段所作的假设称为**原假设**或**零假设**.

为使问题表述得更明确,通常还提出一个与之相对的假设,称为**备择假设**.

原假设与备择假设通常表示为:

$$H_0: \theta \in \Theta_0$$
$$H_1: \theta \in \Theta_1$$

3)双边假设与单边假设

关于一维实参数的假设常有以下形式,其中θ_0为给定的值:

$$\textbf{单边假设} H_0:\theta=\theta_0, H_1:\theta\neq\theta_0.$$

$$\textbf{双边假设}\begin{cases}\text{左边假设} H_0:\theta=\theta_0(\theta\geq\theta_0), H_1:\theta<\theta_0\\ \text{右边假设} H_0:\theta=\theta_0(\theta\leq\ \theta_0), H_1:\theta>\theta_0\end{cases}$$

4)拒绝域/接受域/临界点

在统计检验过程中:

当样本落入**拒绝域**时,拒绝原假设;

当样本落入**接受域**时,接受原假设.

拒绝域的边界点称为**临界点**。

5)显著性水平

对假设检验问题,设$X_1,X_2,\dots,X_n$为样本,W为样本空间的一个子集,对于给定的小概率$\alpha\in(0,1)$,对任意的待推断$\theta\in\Theta_0$ (原假设), 若W满足:

$$P_\theta\left((X_1,X_2,\dots,X_n)\in W\right)\leq\alpha,$$

则W构成了原假设的一个拒绝域..称α为显著性水平,并称此由W构成拒绝域的检验方法为显著性水平为α的检验.

6)假设检验的步骤

根据实际问题所关心的内容, 建立H_0与H_1.

在H_0为真的前提下,选择合适的检验统计量V.

由H_1确定出拒绝域形式,

对于给定的显著性水平α,其对应的拒绝域

$$\begin{cases}\text{双边假设}\left(V<V_{\frac{\alpha}{2}}\right)\cup\left(V>V_{\frac{\alpha}{2}}\right)\\ \text{左边假设}(V<V_\alpha)\\ \text{右边假设}(V>V_{1-\alpha})\end{cases}$$

根据样本值计算v,检查样本是否落入拒绝域W.

得出结论:若落入W,则拒绝H_0,接受H_1;若未落入W,则接受H_0.

→两点注意:

1.α值的选取

(1)必须事先确定.

(2)α值大小选取应依所研究具体问题来决定.

2.对于单侧检验,往往把希望的结果(或预计的结果)的反面取作H_0.

7)两类错误概率

弃真:H_0为真,拒绝H_0

纳伪:H_0为假,接受H_0

理想的检验方法应使犯两类错误的概率都很小.

记第一类错误的概率为α(恰为显著性水平);第二类错误的概率为β

当样本容量一定时,犯两类错误的概率不能同时减小.

一般,作假设检验时,先控制犯第一类错误的概率α,在此基础上使β尽量地小,要降低β一般要增大样本容量.

当H_0不真时,参数值越接近真值,β越大.

H_0与H_1地位本应平等,但在控制犯第一类错误的概率α的原则下,通常将有把握的,有经验的结论作为原假设,或者尽可能使后果严重的错误成为第一类错误.

8)假设检验的评价准则

→功效函数

设θ为总体的待推断参数,对于一个具有拒绝域W的检验τ,定义:

$$\beta(\theta)=P_\theta(W)=P_\theta\left((X_1,X_2,\dots,X_n)\in W\right),\theta\in\Theta$$

为该检验的**功效函数**,也记为$\beta_W(\theta)$或$\beta_\tau(\theta)$.(即为样本落在拒绝域的概率)

$$\alpha=\beta(\theta),\beta=1-\beta(\theta)$$

→一致最优检验$uniformly\ most\ powerful\ test$

对给定的$\alpha\in(0,1)$,设τ^*为一个水平α的检验,若对于任意一个水平α的检验τ,有

$$\beta_{\tau^*}(\theta)\ge\beta_\tau(\theta),\forall\theta\in\Theta_1$$

则称τ^*为一致最优检验(或一致最大功效检验),记为$\boldsymbol{UMP}$**检验.**

→无偏检验$unbiased\ test$

若对任意$\theta_0\in\Theta_0$及$\theta_1\in\Theta_1$,都有:$\beta(\theta_0)\le\beta(\theta_1)$,则称$\tau$为一个**无偏检验**.

即,要求一个检验犯第一类错误的概率总不超过不犯第二类错误的概率.

易得**一致最优无偏检验**的概念

9)正态总体的假设检验(**参数检验**)

$$\text{单正态总体:}\begin{cases}\text{均值}\begin{cases}\sigma^2\text{已知 }U=\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}\sim N(0,1),|U|>u_{1-\alpha/2},U\begin{cases}<u_\alpha\\>u_{1-\alpha}\end{cases}\\ \sigma^2\text{未知}T=\frac{\bar{X}-\mu_0}{S/\sqrt{n}}\sim t(n-1),|T|>t_{1-\alpha/2},T\begin{cases}<-t_{1-\alpha}\\>t_{1-\alpha}\end{cases}\end{cases}\\ \text{方差}\begin{cases}\mu\text{已知}\chi^2=\frac{\sum_{i=1}^n(X_i-\mu)^2}{\sigma_0^2}\sim\chi^2(n),\chi^2\begin{cases}<\chi_\alpha^2(n)\\>\chi_{1-\alpha}^2(n)\end{cases}\\ \mu\text{未知}\chi^2=\frac{(n-1)S^2}{\sigma_0^2}\sim\chi^2(n-1),\chi^2\begin{cases}<\chi_\alpha^2(n-1)\\>\chi_{1-\alpha}^2(n-1)\end{cases}\end{cases}\end{cases}$$

$$\text{双正态总体}\begin{cases}\text{均值差}\begin{cases}\sigma^2\text{已知 }U=\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}}}\sim N(0,1),|U|>u_{1-\alpha/2},U\begin{cases}<u_\alpha\\>u_{1-\alpha}\end{cases}\\ \sigma^2\text{未知},\sigma_1^2=\sigma_2^2,T=\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{1}{n}+\frac{1}{m}}S_W}\sim t(n+m-2),T\begin{cases}<-t_{1-\alpha}\\>t_{1-\alpha}\end{cases}\end{cases}\\ \text{方差比}\quad\mu\text{未知}F=\frac{S_1^2}{S_2^2}\sim F(n-1,m-1),F\begin{cases}<F_\alpha(n-1,m-1)\\>F_{1-\alpha}(n-1,m-1)\end{cases}\end{cases}$$

假设检验与区间估计具有相同的数学结构

10)样本容量的选取$\delta=\mu_1-\mu_0,\lambda=\frac{\mu-\mu_0}{\sigma/\sqrt{n}}$

即给定β,利用β在相应检验方法的计算公式反求n的范围

(对β分别利用定义和分布分位数算两次)

例如,在单正态总体的U检验下,当n满足:

$$\begin{cases}\sqrt{n}\ge\left(u_{1-\alpha}+u_{1-\beta}\right)\sigma/\delta\quad\text{单边检验}\\ \sqrt{n}\ge\left(u_{1-\frac{\alpha}{2}}+u_{1-\beta}\right)\sigma/\delta\quad\text{双边检验}\end{cases}$$

时,犯第二类错误的概率不超过给定的β

→ U检验法中β的计算公式:

$$\begin{cases}\text{右边检验 }\beta=\Phi(u_{1-\alpha}-\lambda)\\ \text{左边检验 }\beta=\Phi(u_{1-\alpha}+\lambda)\\ \text{双边检验 }\beta=\Phi(u_{1-\alpha}-\lambda)+\Phi(u_{1-\alpha}+\lambda)-1\end{cases}$$

11)成对数据均值的检验(利用$Z = X - Y$)实现
12)总体分布的假设检验(**非参数检验**)

拟合优度检验

总体的分布不再是已知属于某种类型的分布族,问题变为对总体是否属于某分布族进行统计推断:基于样本信息,验证总体数据与假定之可能分布的拟合程度。

仍为假设检验问题,但有特点:

可由一个零假设所确定

零假设所确定的分布集合总是比较小的

$$\begin{cases} \chi^2\text{检验法(分布的检验(多项分布))} \\ *\ \text{柯氏检验法(分布的检验(经验分布函数))} \\ \text{偏度/峰度检验法(分布是否为正态的检验)} \\ *\ \text{秩和检验(两连续分布是否相同的检验)} \end{cases}$$

→ $Pearson\chi^2$检验($Karl\ Pearson, 1900$)

设完备事件组$A_1, A_2, \dots, A_k$,原假设$H_0: P(A_i) = p_i, i = 1,2,\dots,k$

进行n次独立重复试验,事件出现的频数分别为$v_1, v_2, \dots, v_k$

$Perason$定理:若假设H_0成立,则当$n \to +\infty$时,

$$V = \sum_{i=1}^{k} \frac{(v_i - np_i)^2}{np_i} \overset{\text{相似}}{\sim} \chi^2(k - r - 1)$$

其中r是用最大似然估计法估计的未知参数的个数.

→拟合优度检验的一般步骤

➢根据实际问题,确定总体可能所属的分布族

$$\mathfrak{F}_0 = \{F(x,\theta): \theta \in \Theta\}$$

➢建立假设$H_0: F \in \mathfrak{F}_0; H_1: F \notin \mathfrak{F}_0$

➢ 在H_0的假定之下,求出分布族参数θ的估计,得到总体可能的具体分布F_0(称为拟合分布)

➢ 修正假设$H_0: F \equiv F_0; H1: F \neq F_0$

➢ 对于给定的显著性水平α,选择适当的检验方法

➢ 确定出其对应的拒绝域W,<u>注意因参数估计所致自由度的变化</u>

➢ 根据样本值计算,检查样本是否落入拒绝域W

➢ 得出结论

→ $Pearson\chi^2$检验是针对离散多项分布的检验

对于连续分布,可作如下推广:

➢首先将实数轴划分成k个区间I_i $(i = 1, \dots, k)$

➢计算拟合分布F_0在各个区间I_i的概率p_i

➢同时计算样本点落在各个区间的频数v_i

➢再利用χ^2检验法

→ 偏度/峰度检验法(样本容量n大于 100 为宜)

对于随机变量X的标准化变量$X^* = \frac{X - E(X)}{\sqrt{D(X)}}$

→若X服从正态分布,则$v_1 = E(X^*)^3$(X的偏度)$= 0, v_2 = E(X^*)^4$(X的峰度)$= 3$

对于$G_1 = B_3 / B_2^{1.5}$ (称为样本偏度),$G_2 = B_4 / B_2^2$ (称为样本峰度),有

$$U_1 = \frac{G_1}{\sigma_1} \sim U_2 = \frac{G_2 - \mu_2}{\sigma_2} \sim N(0,1)$$

其中$\sigma_1^2 = \frac{6(n-2)}{(n+1)(n+3)}, \mu_2 = 3 - \frac{6}{n+1}, \sigma_2^2 = \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}$

拒绝域为$|u_1| \geq u_{1-\frac{\alpha}{4}}$或$|u_2| \geq u_{1-\frac{\alpha}{4}}$

六) 方差分析

试验指标:待考察现象某属性的数量指标

因素(因子):影响指标的条件

可控因素:如:温度\剂量等

不可控因素:如:测量误差\气象条件等

因素水平:因素所处的状态

单因素试验:试验中只有一个因素在起作用

多因素试验:试验中有多个因素在起作用

1)单因素方差分析

→试验次数相等的方差分析:

假定因子A有m个水平,分别记为:$A_1, A_2, \dots A_m$.在每一种水平下,进行k次试验.(即试验次数相等,各水平下试验次数均为k).每次试验可记为X_{ij},表示在第i个水平下的第j次试验.试验完成后可得其观察值:x_{ij}

$$X_{ij} \sim N(\mu_i, \sigma^2), i = 1,2, \dots, m, j = 1,2, \dots, k.$$

设$\varepsilon_i = \mu_i - \mu$是因子$A$的第$i$个水平$A_i$所引起的差异,称为水平$A_i$的**效应**.

其中$\mu = \frac{1}{m}\sum_{i=1}^m \mu_i$称为均值(期望)的总平均,类似地定义$\bar{X}, \bar{X}_i$

总离差平方和:$S_T = \sum_{i=1}^m \sum_{j=1}^{n_i} \left(X_{ij} - \bar{X}\right)^2 = (mk-1)S^2$

→在假设H_0成立时,有$\frac{S_T}{\sigma^2} \sim \chi^2(mk-1)$.

误差平方和(组内平方和):$S_e = \sum_{i=1}^m \sum_{j=1}^{n_i} \left(X_{ij} - \bar{X}_i\right)^2$

效应平方和(组间平方和):$S_A = \sum_{i=1}^m \sum_{j=1}^{n_i} (\bar{X}_i - \bar{X})^2 = k\sum_{i=1}^m (\bar{X}_i - \bar{X})^2$

→ $S_T = S_A + S_e$

→平方和分解定理: 设$Q_1 + Q_2 + \cdots + Q_k = Q \sim \chi^2(n)$

其中Q_i是秩为f_i的非负二次型,则

$$f_1 + f_2 + \cdots + f_k = n \Leftrightarrow Q_i \sim \chi^2(f_i)\text{且相互独立}$$

构造检验统计量:

$$F = \frac{S_A/(m-1)}{S_e/m(k-1)} \sim F(m-1, m(k-1))$$

拒绝域:

$$F_A > F_{1-\alpha}(m-1, m(k-1))$$

→计算公式:

$$T_i=\sum_{j=1}^{n_i}x_{ij},i=1,2,\dots,m;CT=n\bar{x}^2=\frac{1}{n}\left(\sum_{i=1}^{m}T_i\right)^2\text{(称为修正项)}$$

则有:

$$S_T=\sum_{i=1}^{m}\sum_{j=1}^{k}x_{ij}^2-CT,S_A=\sum_{i=1}^{m}\left(\frac{T_i^2}{n_i}\right)-CT$$

→试验次数不等的方差分析(k修正为n_i)

$$n=\sum_{i=1}^{m}n_i,F=\frac{S_A/(m-1)}{S_e/(n-m)}\sim F(m-1,n-m)$$

→方差分析显著性检验具体步骤:

➢数据预处理

$$y_{ij}=\frac{x_{ij}-c}{d}$$

➢列出统计分析数据表

因子水平＼试验次数	1　2　⋯　j　⋯　k	$\sum_{j=1}^{k}x_{ij}$	$\bar{x}_i=\frac{1}{k}\sum_{j=1}^{k}x_{ij}$
A_1	x_{11} $x_{12}\cdots x_{1j}\cdots x_{1k}$	$\sum x_{1j}$	$\bar{x}_1$
A_2	x_{21} $x_{22}\cdots x_{2j}\cdots x_{2k}$	$\sum x_{2j}$	$\bar{x}_2$
⋯	⋯		
A_i	x_{i1} $x_{i2}\cdots x_{ij}\cdots x_{ik}$	$\sum x_{ij}$	$\bar{x}_i$
⋯	⋯		
A_m	x_{m1} $x_{m2}\cdots x_{mj}\cdots x_{mk}$	$\sum x_{mk}$	$\bar{x}_m$

➢按照公式进行计算

➢列出方差分析表 注意试验次数不等引起自由度变化

单因素方差分析表↓

方差来源	平方和	自由度	均方	F比(值)
因素A	S_A	$m-1$	$\bar{S}_A=\frac{S_A}{m-1}$	$F=\frac{\bar{S}_A}{\bar{S}_e}$
误差	S_e	$n-m$	$\bar{S}_e=\frac{S_e}{n-m}$	
总和	S_T	$n-1$		

➢根据显著性水平,进行统计检验

2)双因子方差分析

→双因子间无交互作用的情形

因子A有n个水平:$A_1, A_2, \dots A_n$,因子B有m个水平:$B_1, B_2, \dots, B_m$.
A_iB_j的一次试验,结果以X_{ij}表示,观察值记为x_{ij}.
假定:$X_{ij} \sim N(\mu_{ij}, \sigma^2)$

$$H_0 = H_{01} \cap H_{02}, H_{01}: \alpha_1 = \alpha_2 = \cdots = \alpha_n = 0, H_{02}: \beta_1 = \beta_2 = \cdots = \beta_m = 0$$

$$或H_0: \mu_{111} = \mu_{112} = \cdots = \mu_{nmr}$$

效应可加性:$\mu_{ij} = \mu + \alpha_i + \beta_j$(存在对$\mu_{ij}, \sigma^2$做参数估计的问题)

数据总平均:$\bar{X} = \overline{X_{\cdot\cdot}} = \frac{1}{nm}\sum X_{ij}$

A_i组内平均:$\overline{X_{i\cdot}} = \frac{1}{m}\sum_{j=1}^{m} X_{ij}$, B_j组内平均:$\overline{X_{\cdot j}} = \frac{1}{n}\sum_{i=1}^{n} X_{ij}$

总离差平方和:$S_T = \sum_{i=1}^{n}\sum_{j=1}^{m}\left(X_{ij} - \bar{X}\right)^2$

A效应平方和:$S_A = m\sum_{i=1}^{n}(\overline{X_{i\cdot}} - \bar{X})^2$, B效应平方和:$S_B = n\sum_{j=1}^{m}\left(\overline{X_{\cdot j}} - \bar{X}\right)^2$

误差平方和:$S_e = \sum_{i=1}^{n}\sum_{j=1}^{m}\left(X_{ij} - \overline{X_{i\cdot}} - \overline{X_{\cdot j}} + \bar{X}\right)^2$

$S_T = S_A + S_B + S_e$

构造检验统计量:$F_A = \frac{\frac{S_A}{n-1}}{\frac{S_e}{(n-1)(m-1)}}$, $F_B = \frac{\frac{S_B}{n-1}}{\frac{S_e}{(n-1)(m-1)}}$

方便计算:$T_{i\cdot}, T_{\cdot j}, T_{\cdot\cdot}, TS = \sum_{i=1}^{n}\sum_{j=1}^{m} X_{ij}^2$

$$S_A = \frac{1}{m}\sum T_{i\cdot}^2 - \frac{T_{\cdot\cdot}^2}{nm}, S_B = \frac{1}{n}\sum T_{\cdot j}^2 - \frac{T_{\cdot\cdot}^2}{nm}, S_T = TS - \frac{T_{\cdot\cdot}^2}{nm}$$

→双因子间有交互作用的情形

因子A有n个水平:$A_1, A_2, \dots A_n$,因子B有m个水平:$B_1, B_2, \dots, B_m$每个水平r次实验
A_iB_j的第k次试验,结果以X_{ijk}表示,观察值记为x_{ijk}.
假定:$X_{ijk} \sim N(\mu_{ijk}, \sigma^2)$

$$H_0 = H_{01} \cap H_{02} \cap H_{03}, H_{01}: \alpha_1 = \alpha_2 = \cdots = \alpha_n = 0,$$
$$H_{02}: \beta_1 = \beta_2 = \cdots = \beta_m = 0, H_{03}: \gamma_{11} = \gamma_{12} = \cdots = \gamma_{nm} = 0$$

$$或H_0: \mu_{111} = \mu_{112} = \cdots = \mu_{nmr}$$

效应可加性:$\mu_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij}$($\gamma_{ij}$表示$A_i, B_j$的交互效应)

数据总平均:$\bar{X} = \overline{X_{\cdot\cdot\cdot}} = \frac{1}{nmr}\sum X_{ijk}$

A_i组内平均:$\overline{X_{i\cdot\cdot}} = \frac{1}{m}\sum_{j=1}^{m}\overline{X_{ij\cdot}}$, B_j组内平均:$\overline{X_{\cdot j\cdot}} = \frac{1}{n}\sum_{i=1}^{n}\overline{X_{ij\cdot}}$

A_iB_j交互组内平均值:$\overline{X_{ij\cdot}} = \frac{1}{r}\sum_{k=1}^{r} X_{ijk}$

总离差平方和:$S_T = \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{r}\left(X_{ijk} - \bar{X}\right)^2$

A效应平方和:$S_A = mr\sum_{i=1}^{n}(\overline{X_{i\cdot\cdot}} - \bar{X})^2$, B效应平方和:$S_B = nr\sum_{j=1}^{m}\left(\overline{X_{\cdot j\cdot}} - \bar{X}\right)^2$

交互效应平方和$S_{AB}=r\sum_{i=1}^{n}\sum_{j=1}^{m}\left(\overline{X_{ij\cdot}}-\overline{X_{i\cdot\cdot}}-\overline{X_{\cdot j\cdot}}+\bar{X}\right)^2$

误差平方和:$S_e=\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{r}\left(X_{ijk}-\overline{X_{ij\cdot}}\right)^2$

$S_T=S_A+S_B+S_{AB}+S_e$

构造检验统计量:$F_A=\frac{\frac{S_A}{n-1}}{\frac{S_e}{mn(r-1)}},F_B=\frac{\frac{S_B}{n-1}}{\frac{S_e}{mn(r-1)}},F_{AB}=\frac{\frac{S_{AB}}{(n-1)(m-1)}}{\frac{S_e}{mn(r-1)}}$

方便计算:$T_{i\cdot\cdot},T_{\cdot j\cdot},T_{\cdots},TS=\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{r}X_{ijk}^2$

$$S_A=\frac{1}{mr}\sum T_{i\cdot\cdot}^2-\frac{T_{\cdots}^2}{nmr},S_B=\frac{1}{nr}\sum T_{\cdot j\cdot}^2-\frac{T_{\cdots}^2}{nmr},S_T=TS-\frac{T_{\cdots}^2}{nmr}$$

七) 回归分析

1)一元线性回归

样本点$(x_i,y_i),i=1,2,\dots,n$

线性相关关系$Y=a+bx+\varepsilon,\varepsilon\sim N(0,\sigma^2)$

对回归系数的估计值$\hat{a},\hat{b}$,可取$\hat{y}=\hat{a}+\hat{b}x$,作为其相应Y之观测值y的估计

→极大似然估计法

引入统计量:$S_{xy}=\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})=\sum_{i=1}^{n}x_iy_i-\frac{1}{n}(\sum_{i=1}^{n}x_i)(\sum_{i=1}^{n}y_i)$

求解$Y_i\sim N(a+bx_i,\sigma^2)$对应似然函数$L(a,b)$极大值取值得:

$$\hat{b}=\frac{S_{xy}}{S_{xx}},\hat{a}=\frac{\sum y}{n}-\frac{\sum x}{n}\hat{b}$$

→最小二乘法

给定回归系数估计后,存在误差$\delta_i=y_i-\hat{y}_i$,记误差平方和为:$Q=\sum_{i=1}^{n}\delta_i^2$

求其极小值取值,发现其估计与极大似然估计法等价,且$\hat{a},\hat{b},\hat{y}$无偏

2)回归方程的显著性检验

→相关系数检验法

样本相关系数$|r|=\sqrt{1-\frac{Q}{S_{yy}}}$

检验统计量$t=\frac{\sqrt{n-2}r}{\sqrt{1-r^2}}\sim t(n-2),|t|>t_{1-\frac{\alpha}{2}}(n-2)$时拒绝原假设,存在线性关系

→ F检验法:假设$H_0:b=0$

检验统计量$F=\frac{(n-2)U}{Q}\sim F(1,n-2),F>F_{1-\alpha}$时拒绝原假设,存在线性关系

其中$U=S_{yy}-Q=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2=\frac{S_{xy}^2}{S_{xx}}$

3)回归分析的基本步骤(一元线性回归)

➢线性变换预处理(方便计算)

➢求解回归系数的估计,得到回归方程

➢在给定显著性水平下进行统计检验

➢利用可信回归直线进行预测或控制

4)多元线性回归

$Y=b_0+b_1x_1+b_2x_2+\cdots+b_kx_k+\varepsilon,\varepsilon\sim N(0,\sigma^2)$,观察值$(x_{1t},x_{2t},\dots,x_{kt};y_t),1\le t\le n$

最小二乘法:$\frac{\partial Q}{\partial b_i}=0$

5)可化成线性回归的非线性回归问题

$y = b_0 + \sum_{j=1}^{k} b_j g_j(x_1, x_2, \dots, x_m) + \varepsilon$(注意维数变化,有$m$个变元但是是$k$元回归)

逐步回归:从众多的自变量中,根据这些变量各自对回归方程影响的大小,逐次地选入到回归方程中,并将那些由于新变量引入而失去重要性的变量在回归方程中淘汰,持续上述步骤,直至回归方程不再有可淘汰的变量且没有再可引入的变量时为止.