1 概率论基础

- (蒲丰投针问题, 2020.2.25)桌面上画满间隔为a的平行线, 投长为$l(l<a)$的针, 求事件 $E=\{$针与某直线相交$\}$的概率.

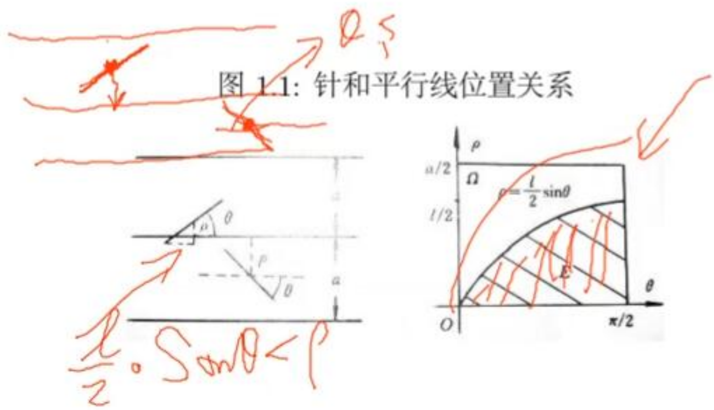
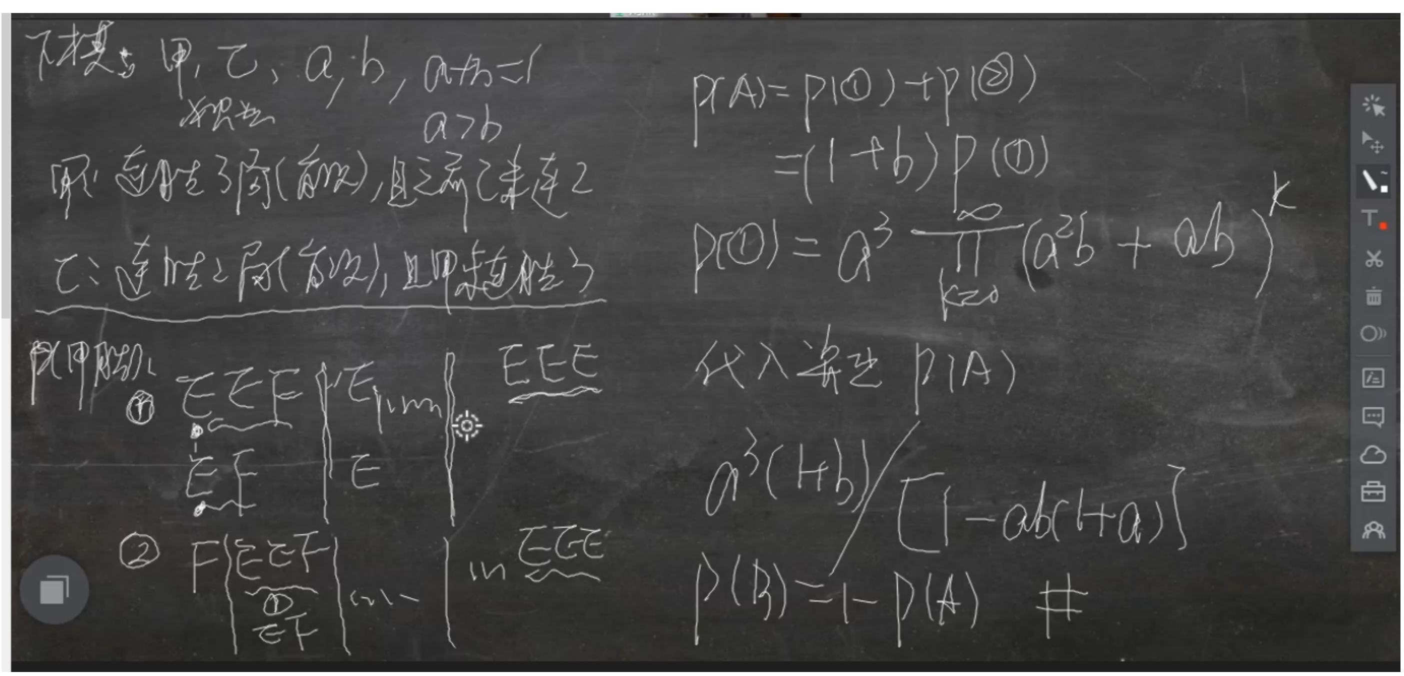


图 1: a

- (Polya罐子模型, 2020.2.25)罐子中有a个白球b个黑球, 每次摸出一个后连同c个同色球放回. 证明第n次取球, 取出白球概率为$\frac{a}{a+b}$

假设第$n=k-1$次取, 概率为$\frac{a}{a+b}$,
则有$P(A_k|A_1)=\frac{(a+c)}{(a+c)+b}$, $P(A_k|\overline{A_1})=\frac{(a)}{a+(b+c)}$,
(实际上是假设最开始有a+c个白球, b个黑球, 这样$P(A_k)$)当然就是$\frac{(a+c)}{(a+c)+b}$了)
因此$P(A_k)=P(A_1)P(A_k|A_1)+P(\overline{A_1})P(A_k|\overline{A_1})=\frac{a}{a+b}$



2 随机变量及其分布

- 离散

分布	标记	公式	均值	方差	特点
0-1分布		$P(X=1)=p, P(X=0)=1-p$	p	0	
伯努利分布	$Ber(p)$		p	$p(1-p)$	
二项分布	$B(n,p)$	$P(X=k)=\begin{pmatrix} n \\ k \end{pmatrix} p^k(1-p)^{n-k}$	np	$np(1-p)$	极限是泊松分布 再生性
几何分布	$G(p)$	$P(X=k)=(1-p)^{k-1}p$	$\frac{1}{p}$	$\frac{q}{p^2}$	无记忆性
Poisson分布	$P(\lambda)$	$P(X=k)=\frac{\lambda^k}{k!}e^{-\lambda}, \lambda>0$	λ	λ	再生性
离散均匀分布	$Ua_1, a_2, \cdots, a_n$				

- 连续

分布	标记	公式	均值	方差	特点
正态分布	$N(\mu,\sigma)$	$f(x)=\frac{1}{\sqrt{2\pi}}\sigma\exp-\frac{(x-\mu)^2}{2\sigma^2}$	μ	σ^2	$aX+bY\sim N(a\mu_1+b\mu_2, a^2\sigma_1^2+b^2\sigma_2^2)$
指数分布	$Exp(\lambda)$	$f(x)=\begin{cases} \lambda e^{-\lambda x}, x>0 \\ 0, x\le 0 \end{cases}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	无记忆性
Weibull分布		$f(x)=\begin{cases} \lambda\alpha x^{\alpha-1}e^{-\lambda x^{\alpha}}, x>0, \alpha>0 \\ 0, x\le 0 \end{cases}$			
均匀分布		$f(x)=\begin{cases} \frac{1}{b-a}, a\le x\le b \\ 0, else \end{cases}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	

- 多维

分布	标记	公式	均值	方差	特点
多项分布		$P(X_1=k_1,\cdots,X_n=k_n)=\frac{n!}{k_1!\cdots k_n!}p_1^k\cdots p_n^k$			$(p_1+\cdots+p_n)^n=$
二元正态	$N(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$	$\frac{1}{(2\pi\sigma_1\sigma_2\sqrt{1-\rho^2})}$ $\exp\left\{-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu_1)^2}{\sigma_1^2}-\frac{2\rho(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2}+\frac{(y-\mu_2)^2}{\sigma_2^2}\right)\right\}$			

- $\Phi(x)$和$\phi(x)$表示$N(0,1)$的分布函数和密度函数
- 指数函数通常用来描述失效率(无老化寿命分布), 即在Δx这段时间内, 失效概率恒为λ
- 失效率函数$\frac{F'(x)}{1-F(x)}=\frac{P(x\le X\le x+\delta x|X>x)}{\delta x}$
- 三大分布
  - 卡方分布: $X\sim\chi_n^2$, $X=\sum\limits_{i=1}^{n}X_i^2$
    - $EX=n$, $Var(X)=2n$
  - t分布: $T\sim t_n$, $T=\frac{X_1}{\sqrt{X_2/n}}$,$X_1\sim N(0,1)$, $X_2\sim\chi_n^2$
    - $ET=0$当$n\ge 2$, $Var(T)=\frac{n}{n-2}$当$n\ge 3$
    - 类似标准正态, n无穷时趋于标准正态
  - F分布: $F\sim F_{n,m}$, $X_1\sim\chi_n^2, X_2\sim\chi_m^2, F=\frac{X_1/n}{X_2/m}$
    - 图像类似卡方分布
    - 若$Z\sim F_{m,n}$, 则$\frac{1}{Z}\sim F_{n,m}$

* 若$T \sim t_n$, 则$T^2 \sim F_{1,n}$
* $F_{m,n}(1-\alpha) = \frac{1}{F_{n,m}(\alpha)}$

2.1 连续型随机变量的条件分布

有

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

2.2 随机变量的函数的概率分布

$y = g(x)$严格单调连续, 反函数唯一为$x = h(y)$, 且$h'(y)$存在且连续, 则$Y = g(x)$也是连续型随机变量, 有

$$p(y) = f(h(y))|h'(y)|$$

多元的则乘上Jacobi行列式的绝对值即可.

3 数字特征

3.1 期望

- $E(aX + bY) = aE(X) + b(Y)$
- 独立随机变量$E(XY) = E(X)E(Y)$
- $Eg(X) = \begin{cases} \sum g(a_i)p_i \\ \int_{-\infty}^{\infty} g(x)f(x)dx \end{cases}$
- $E(Y|X = x) = \begin{cases} \sum a_i p_i \\ \int_{-\infty}^{\infty} yf(y|x)dy \end{cases}$
- 全期望公式: $E(E(Y|X)) = E(g(X)) = EY$
- p分位数: $P(X \le \mu_p) \ge p$则μ_p是随机变量X的p分位数

3.2 方差

- $Var(X) = E(X - EX)^2 = \sigma^2 = EX^2 - (EX)^2$
- 若X,Y不相关, 则$Var(aX + bY) = a^2 Var(X) + b^2 Var(Y)$
- 标准化随机变量: $X^* = \frac{X - EX}{\sqrt{Var(X)}}, EX^* = 0, Var(X^*) = 1$

3.3 矩(Moment)

- 矩: $E[(X - c)^r]$
- 原点矩: $c = 0$, 即EX^r
- 中心矩: $E[(X - EX)^r]$

- 均值是一阶原点矩, 方差是二阶中心矩
- k阶阶乘矩: $E[X(X-1)\cdots(X-k+1)]$

3.4 协方差

- $Cov(X,Y)=E(X-EX)(Y-EY)=EXY-EXEY$
- $Var(X+Y)=Var(X)+Var(Y)+2E(X-EX)(Y-EY)$
- 独立则$Cov(X,Y)=0$
- $Cov(a_1X_1+a_2X_2,b_1Y_1+b_2Y_2)=\sum\limits_{i=1}^{2}\sum\limits_{j=1}^{2}a_ib_jCov(X_i,Y_j)$
- 二元正态分布$(X,Y)\sim N(a,b,\sigma_1^2,\sigma_2^2,\rho)$, 则协方差矩阵为$\begin{bmatrix}\sigma_1^2 & \rho\sigma_1\sigma 2\\ \rho\sigma_1\sigma 2 & \sigma_2^2\end{bmatrix}$
- $[Cov(X,Y)]^2\leq\sigma_1^2\sigma_2^2=Var(X)Var(Y)$等号成立当且仅当$Y=aX+b$

3.5 相关系数

- $\rho_{X,Y}=\frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}\in[-1,1]$, 等号成立当且仅当$X,Y$之间存在严格线性关系, $\rho_{X,Y}$的正负与线性关系斜率正负相同
- 衡量变量之间的线性强度
- $\rho=0$时称X,Y不相关, 只表示无线性关系
- ξ,η不相关$\Leftrightarrow Cov(\xi,\eta)=0\Leftrightarrow E\xi\eta=E\xi E\eta\Leftrightarrow Var(\xi+\eta)=Var(\xi)+Var(\eta)$
- 独立一定不相关, 不相关却不一定独立. 只在二元正态分布下, 独立$\Leftrightarrow$不相关

4 大数定律和中心极限定理

4.1 大数定律

- 弱大数定律: 对任何$\epsilon>0$, $\lim\limits_{n\to\infty}P(|\xi_n-\xi|\geq\epsilon)=0$
- 独立同分布数列$\{X_n\}$服从大数定律, $\bar{X}=\frac{1}{n}\sum\limits_{k=1}^{n}X_k\xrightarrow{P}\mu$, μ为X_n期望. 即依概率收敛到X_n的期望μ
- Markov不等式: Y非负, $\forall\epsilon>0,P(Y\geq\epsilon)\leq\frac{EY}{\epsilon}$
- 切比雪夫不等式: $P(|X-EX|\geq\epsilon)\leq\frac{Var(X)}{\epsilon^2}$

4.2 中心极限定理

- $\{X_n\}$独立同分布, 期望方差为μ,σ^2, 则$\sum\limits_{i=1}^{n}X_i$的标准化形式$\frac{\sum_{i=1}^{n}X_i-n\mu}{\sqrt{n}\sigma}$满足中心极限定理: $\lim\limits_{n\to\infty}F_n(x)=\Psi(x)$, $F_n(x)$为标准化形式的分布函数. 即$\frac{\sum_{i=1}^{n}X_i-n\mu}{\sqrt{n}\sigma}\xrightarrow{d}N(0,1)$(依分布收敛)

5 数理统计的基本概念

5.1 常用统计量

- 样本均值: $\bar{X} = \frac{1}{n}\sum\limits_{i=1}^{n} X_i$
- 样本方差：$S^2 = \frac{1}{n-1}\sum\limits_{i=1}^{n}(X_i - \bar{X})^2$
- 样本原点矩: $a_k = \frac{1}{n}\sum\limits_{i=1}^{n} X_i^k$
- 样本中心矩: $m_k = \frac{1}{n}\sum\limits_{i=1}^{n}(X_i - \bar{X})^k$. 如$\frac{n-1}{n}S^2 = m_2$
- 次序统计量: 排好序的$X_1 \le X_2 \le \cdots \le X_n$, 则$(X_1, X_2, \cdots, X_n)$称为次序统计量
  - 样本中位数: $m_{\frac{1}{2}} = \begin{cases} X_{(\frac{n+1}{2})} & \text{n is odd} \\ \frac{1}{2}\left[X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}\right] & \text{n is even} \end{cases}$
  - 极值
- 经验分布函数

5.2 重要定理

1. 定理1: $X_i\ i.i.d. \sim N(a, \sigma^2)$, $X = \frac{1}{n}\sum\limits_{i=1}^{n} X_i$和$S^2 = \frac{1}{n-1}\sum\limits_{i=1}^{n}(X_i - \bar{X})^2$分别为样本均值和样本方差, 则有
   (a) $\bar{X} \sim N(a, \frac{1}{n}\sigma^2)$
   (b) $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$
   (c) $\bar{X}$和S^2独立
2. $X_i\ i.i.d. \sim N(a, \sigma^2)$,
$$T = \frac{\sqrt{n}(\bar{X} - a)}{S} \sim t_{n-1}$$
3. $X_i\ i.i.d. \sim N(a_1, \sigma_1^2)$, $Y_i\ i.i.d. \sim N(a_2, \sigma_2^2)$, 且$\sigma_1^2 = \sigma_2^2 = \sigma^2$, 相互独立. 则
$$T = \frac{(\bar{X} - \bar{Y}) - (a_1 - a_2)}{S_\omega}\sqrt{\frac{mn}{m+n}} \sim t_{n+m-2}$$
   这里$(n+m-2)S_\omega = (m-1)S_1^2 + (n-1)S_2^2$. 将$(\bar{X} - \bar{Y})$标准化
4. $X_i\ i.i.d. \sim N(a_1, \sigma_1^2)$, $Y_i\ i.i.d. \sim N(a_2, \sigma_2^2)$,
$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{m-1,n-1}$$
5. $X_i\ i.i.d. \sim Exp(\lambda)$, 则
$$2\lambda n\bar{X} = 2\lambda\sum_{i=1}^{n} X_i \sim \chi^2_{2n}$$
6. 三大分布

- 卡方分布: $X \sim \chi_n^2$, $X = \sum\limits_{i=1}^{n} X_i^2$
  - $EX = n$, $Var(X) = 2n$
- t分布: $T \sim t_n$, $T = \frac{X_1}{\sqrt{X_2/n}}$,$X_1 \sim N(0,1)$, $X_2 \sim \chi_n^2$
  - $ET = 0$当$n \geq 2$, $Var(T) = \frac{n}{n-2}$当$n \geq 3$
  - 类似标准正态, n无穷时趋于标准正态
- F分布: $F \sim F_{n,m}$, $X_1 \sim \chi_n^2, X_2 \sim \chi_m^2, F = \frac{X_1/n}{X_2/m}$
  - 图像类似卡方分布
  - 若$Z \sim F_{m,n}$, 则$\frac{1}{Z} \sim F_{n,m}$
  - 若$T \sim t_n$, 则$T^2 \sim F_{1,n}$
  - $F_{m,n}(1-\alpha) = \frac{1}{F_{n,m}(\alpha)}$

6 参数估计

6.1 点估计

6.1.1 矩估计

6.1.2 最大似然估计

6.1.3 点估计的优良准则

- 弱相合估计: $\lim\limits_{n \to \infty} P_{\theta_1,\cdots,\theta_n}(|T(X_1,\cdots,X_n) - g(\theta_1,\cdots,\theta_n)| \geq \epsilon) = 0$
- 无偏性: 设$\hat{g}(X_1,\cdots,X_n)$为待估计参数$g(\theta)$的一个估计量, 若

$$E\hat{g}(X_1,\cdots,X_n) = g(\theta)$$

  则称$\hat{g}(X_1,\cdots,X_n)$为$g(\theta)$的**无偏估计量**
- 相对有效性: 比较两个估计的方差, 小的更有效

6.2 区间估计

6.2.1 枢轴变量法

1. 单正态总体. $X_1, X_2, \cdots, X_n i.i.d. N(\mu, \sigma^2)$
   - 估计μ, 未知σ

$$\frac{\sqrt{n}(\overline{X} - \mu)}{S} \sim t_{n-1}$$

   - 估计σ, 未知μ

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$$

   - 估计σ, 已知$\mu = \mu_0$.

$$\sum_{i=0}^{n} \frac{(X_i - \mu_0)^2}{\sigma^2} \sim \chi_n^2$$

- 估计μ, 已知$\sigma = \sigma_0$ 由$\overline{X} \sim N(\mu, \frac{1}{n}\sigma_0^2)$ 有
$$\frac{\overline{X} - \mu}{\sqrt{\frac{1}{n}\sigma_0^2}} \sim N(0,1)$$

2. 二正态总体. $X_1, X_2, \cdots, X_m i.i.d. N(\mu_1, \sigma_1^2)$, $Y_1, Y_2, \cdots, Y_n i.i.d. N(\mu_2, \sigma_2^2)$, 两组样本之间相互独立

- 估计$\mu_1 - \mu_2$, 未知σ_1, σ_2.
根据$(\overline{X} - \overline{Y}) \sim N(\mu_1 - \mu_2, \frac{1}{n}\sigma_1^2 + \frac{1}{m}\sigma_2^2)$有
$$\frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{S_\omega \sqrt{\frac{m+n}{mn}}} \sim t_{n+m-2}$$
这里$S_\omega = \frac{1}{m+n-2}(\sum_{i=0}^{m}(X_i - \overline{X})^2 + \sum_{i=0}^{n}(Y_i - \overline{Y}))^2$
- 估计$\frac{\sigma_1}{\sigma_2}$, 未知$\mu_1, \mu_2$.
根据$\frac{(n-1)S_1^2}{\sigma_1^2} \sim \chi_{n-1}^2$
$$\frac{S_1^2}{S_2^2}\frac{\sigma_2^2}{\sigma_1^2} \sim F_{n-1,m-1}$$
这里计算时注意$F_{n-1,m-1}(1 - \frac{\alpha}{2}) = 1/F_{m-1,n-1}(\frac{\alpha}{2})$
- 估计$\mu_1 - \mu_2$, 已知$\sigma_1 = \sigma_2 = \sigma_0$. 根据$(\overline{X} - \overline{Y}) \sim N(\mu_1 - \mu_2, \frac{1}{n}\sigma_1^2 + \frac{1}{m}\sigma_2^2)$有
$$\frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sigma_0 \sqrt{\frac{m+n}{mn}}} \sim N(0,1)$$
- 估计$\frac{\sigma_1}{\sigma_2}$, 已知$\mu_1, \mu_2$.
根据$\sum_{i=0}^{n} \frac{(X_i - \mu_1)^2}{\sigma_1^2} \sim \chi_n^2$及$\sum_{i=0}^{m} \frac{(Y_i - \mu_2)^2}{\sigma_2^2} \sim \chi_m^2$有
$$\frac{\frac{1}{n}\sum_{i=0}^{n}(X_i - \mu_1)^2}{\frac{1}{m}\sum_{i=0}^{m}(Y_i - \mu_2)^2} \cdot \frac{\sigma_2^2}{\sigma_1^2} \sim F_{n,m}$$
这里计算时注意$F_{n,m}(1 - \frac{\alpha}{2}) = 1/F_{m,n}(\frac{\alpha}{2})$

6.2.2 大样本法

利用中心极限定理即可

7 假设检验

7.1 一样本正态总体

表 7.2.1: 一样本正态总体$N(\mu, \sigma^2)$.

检验对象	检验统计量	分布	拒绝域†
μ (σ^2已知)	$Z = \sqrt{n}(\bar{X} - \mu_0)/\sigma$	$N(0,1)$	$\begin{cases} \|Z\| > u_{\alpha/2} \\ Z > u_\alpha \\ Z < -u_\alpha \end{cases}$
μ (σ^2 未知)	$T = \sqrt{n}(\bar{X} - \mu_0)/S$	t_{n-1}	$\begin{cases} \|T\| > t_{n-1}(\alpha/2) \\ T > t_{n-1}(\alpha) \\ T < -t_{n-1}(\alpha) \end{cases}$
σ^2 (μ 已知)	$\chi^2 = \frac{1}{\sigma_0^2}\sum_{i=1}^n (X_i - \mu)^2$	χ_n^2	$\begin{cases} \chi^2 > \chi_n^2(\alpha/2)\text{或者}\chi^2 < \chi_n^2(1-\alpha/2) \\ \chi^2 > \chi_n^2(\alpha) \\ \chi^2 < \chi_n^2(1-\alpha) \end{cases}$
σ^2 (μ 未知)	$\chi^2 = \frac{1}{\sigma_0^2}\sum_{i=1}^n (X_i - \bar{X})^2$	χ_{n-1}^2	$\begin{cases} \chi^2 > \chi_{n-1}^2(\alpha/2)\text{或者}\chi^2 < \chi_{n-1}^2(1-\alpha/2) \\ \chi^2 > \chi_{n-1}^2(\alpha) \\ \chi^2 < \chi_{n-1}^2(1-\alpha) \end{cases}$

†有关均值的检验: 对立假设分别为$\mu \neq \mu_0$, $\mu > \mu_0$ 和$\mu < \mu_0$. 有关方差的检验: 对立假设分别为$\sigma^2 \neq \sigma_0^2$, $\sigma^2 > \sigma_0^2$ 和$\sigma^2 < \sigma_0^2$.

7.2 两样本正态总体

表 7.2.2: 两样本正态总体的假设检验

检验对象	检验统计量	分布	拒绝域†
均值(方差已知)	$Z=\frac{\bar{X}-\bar{Y}}{\sqrt{\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}}}$	$N(0,1)$	$\begin{cases}\lvert Z\rvert>u(\alpha/2)\\ Z>u(\alpha)\\ Z<-u(\alpha)\end{cases}$
均值(方差未知)‡	$T=\frac{\bar{X}-\bar{Y}}{S_w\sqrt{\frac{1}{m}+\frac{1}{n}}}$	t_{m+n-2}	$\begin{cases}\lvert T\rvert>t_{m+n-2}(\alpha/2)\\ T>t_{m+n-2}(\alpha)\\ T<-t_{m+n-2}(\alpha)\end{cases}$
方差(均值已知)	$F=\frac{\sum_{i=1}^m(X_i-\mu_1)^2/m}{\sum_{i=1}^n(X_i-\mu_2)^2/n}$	$F_{m,n}$	$\begin{cases}F>F_{m,n}(\alpha/2)\text{或}F<\frac{1}{F_{n,m}(\alpha/2)}\\ F>F_{m,n}(\alpha)\\ F<\frac{1}{F_{n,m}(\alpha)}\end{cases}$
方差(均值未知)	$F=\frac{S_1^2}{S_2^2}$	$F_{m-1,n-1}$	$\begin{cases}F>F_{m-1,n-1}(\alpha/2)\text{或}F<\frac{1}{F_{n-1,m-1}(\alpha/2)}\\ F>F_{m-1,n-1}(\alpha)\\ F<\frac{1}{F_{n-1,m-1}(\alpha)}\end{cases}$

†有关均值的检验: 对立假设分别为$\mu_1\neq\mu_2$, $\mu_1>\mu_2$ 和$\mu_1<\mu_2$. 有关方差的检验: 对立假设分别为$\sigma_1^2\neq\sigma_2^2$, $\sigma_1^2>\sigma_2^2$ 和$\sigma_1^2<\sigma_2^2$.
‡假定方差相等

7.3 成对数据

作差构造一样本正态总体

7.4 0-1分布参数p的检验

由中心极限定理构造统计量$T=\sqrt{n}\frac{\bar{X}-p}{\sqrt{p_0(1-p_0)}}$

7.5 拟合优度检验

判断是不是来自某个分布. H_0: X服从分布F. 一般不必再写一个对立假设.
用Pearson提出的卡方拟合优度检验.

7.5.1 离散

理论频数和实际观测频数. 用

$$T=\sum_{i=1}^{k}\frac{(n_i-np_i)^2}{np_i}\sim\chi_{k-1}^2$$

这里是将统计量n_i分布看成泊松分布, 均值方差都为np_i, 再由中心极限定理得到的, 其平方求和则为卡方分布. 但$\sum\limits_{i=1}^{k}n_i=k$, 因此自由度应该是$k-1$.

拒绝域: $T > \chi^2_{k-1}(\alpha)$

若含有r个未知参数, 要用最大似然估计代替这些参数, 则得到统计量

$$T = \sum_{i=1}^{k} \frac{(n_i - n\hat{p}_i)^2}{n\hat{p}_i} \sim \chi^2_{k-1-r}$$

7.5.2 列联表的独立性和齐一性检验

- 独立性: H_0: 属性A与属性B独立. 用列联表.
- 齐一性: 检验生存和死亡和住哪个医院无关. 转化为独立性. A的各个水平对应的B属性分布一致.

7.5.3 连续

分区间转化为离散型. 一般要求$n\hat{p}_i \geq 5$, 如果不满足需要合并.