数据统计期末背记清单

Made by: X.B.D.

一、基础知识与重要分布

① 三大分布

·卡方分布 $\chi^2(n)$

△$E(X) =$ n　　　$D(X) =$ $2n$

△X与Y相互独立，$X \sim \chi^2(m)$，$Y \sim \chi^2(n)$，则 $X+Y \sim$ $\chi^2(m+n)$

要 ·T分布

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

其中分子表示 $X \sim N(0,1)$，分母表示 $Y \sim \chi^2(n)$

$$\frac{\sum_{i=1}^{n}(x_i-\mu_0)^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = 1+\frac{T^2}{n-1}$$

内部需包含一个自由度为n-1的t分布T的函数u(T)

其中 $T = \dfrac{\sqrt{n(n-1)}(\bar{x}-\mu_0)}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}} \sim t(n-1)$

·F分布

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

其中分子表示 $X \sim \chi^2(m)$，分母表示 $Y \sim \chi^2(n)$

② 常见分布与不常见分布

	EX	DX	表达式
(0,1)分布	p	$p(1-p)$	$p^x(1-p)^{1-x}$
几何分布	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$(1-p)^{k-1}p$
二项分布B(n,p)	np	$np(1-p)$	略
均匀分布U(a,b)	$\frac{a+b}{2}$	$\frac{(a-b)^2}{12}$	×
指数分布Exp(λ)	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\lambda e^{-\lambda x}$
正态分布	μ	σ^2	$\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
泊松分布P(λ)	λ	λ	$\frac{\lambda^k}{k!}e^{-\lambda}$
伽马分布Γ(α,β)	$\frac{\alpha}{\beta}$	α/β^2	$\frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}$
逆伽马分布IG(α,β)	$\frac{\beta}{\alpha-1}$ (当$\alpha>1$)	$\frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$	$\frac{\beta^\alpha}{\Gamma(\alpha)}x^{-\alpha-1}e^{-\frac{\beta}{x}}$

	EX	DX	表达式
贝塔分布β(a,b)	$\frac{a}{a+b}$	$\frac{ab}{(a+b)^2(a+b+1)}$	$\frac{1}{B(a,b)}x^{a-1}(1-x)^{b-1}$

伽马函数：$\Gamma(x) = \int_0^{+\infty} t^{x-1}e^{-t}dt$

$\Gamma(x+1) =$ x $\Gamma(x)$，$\Gamma(1) =$ 1

贝塔函数：$B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx$

$B(\alpha,\beta) = \dfrac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ (用Γ函数表示)

n个泊松分布$P(\lambda)$的和为 $P(n\lambda)$

n个指数分布$Exp(\lambda)$的和为 $\Gamma(n,\lambda)$

$X \sim \Gamma(n,\lambda)$，则 $aX \sim \Gamma(n, \frac{\lambda}{a})$ 首猫不忍 厚徳乱

$\Gamma(n,\frac{1}{2}) =$ $\chi^2(2n)$

二、枢轴量与检验统计量

① 枢轴量

未知参数	条件	枢轴量及分布
μ	σ^2已知	$(\bar X-\mu)/(\sigma/\sqrt n)\sim N(0,1)$
	σ^2未知	$(\bar X-\mu)/(S/\sqrt n)\sim t(n-1)$
$\mu_1-\mu_2$	σ_1^2、σ_2^2已知	$(\bar X-\bar Y)-(\mu_1-\mu_2)/\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\sim N(0,1)$
	$\sigma_1^2=\sigma_2^2$未知	$(\bar X-\bar Y)-(\mu_1-\mu_2)/S_w\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\sim t(n_1+n_2-2)$ ← 此处的 $S_w^2=\dfrac{(m-1)S_1^2+(n-1)S_2^2}{m+n-2}$
σ^2	μ已知	$\sum_{i=1}^n(X_i-\mu)^2/\sigma^2\sim\chi^2(n)$
	μ未知	$(n-1)S^2/\sigma^2\sim\chi^2(n-1)$
σ_1^2/σ_2^2	μ_1、μ_2未知	$\dfrac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2}\sim F(n_1-1,n_2-1)$

② 检验统计量（似然比）

	条件 $X\sim N(\mu,\sigma^2)$ μ、σ^2未知	否定域及属于哪类分布	
T_0类	$H_0:\mu=\mu_0\leftrightarrow H_1:\mu\neq\mu_0$	$\left\{\lvert T\rvert=\left\lvert\dfrac{\sqrt{n(n-1)}\,(\bar X-\mu_0)}{\sqrt{\sum_{i=1}^n(X_i-\bar X)^2}}\right\rvert>c\right\}$	$T\sim t(n-1)$
	$H_0:\mu\leq\mu_0\leftrightarrow H_1:\mu>\mu_0$	$\left\{T=\dfrac{\sqrt{n(n-1)}\,(\bar X-\mu_0)}{\sqrt{\sum_{i=1}^n(X_i-\bar X)^2}}>c\right\}$	$T\sim t(n-1)$
T_1类	$H_0:\sigma^2=\sigma_0^2\leftrightarrow H_1:\sigma^2\neq\sigma_0^2$	$\left\{T_1=\dfrac{\sum_{i=1}^n(X_i-\bar X)^2}{\sigma_0^2}>c_2\text{ 或 }<c_1\right\}$	$T_1\sim\chi^2(n-1)$
	$H_0:\sigma^2\leq\sigma_0^2\leftrightarrow H_1:\sigma^2>\sigma_0^2$	$\left\{T_1=\dfrac{\sum_{i=1}^n(X_i-\bar X)^2}{\sigma_0^2}>c\right\}$	$T_1\sim\chi^2(n-1)$

	条件	$X\sim N(\mu_1,\sigma_1^2)$ $X_1,X_2\cdots X_{n_1}$ $Y\sim N(\mu_2,\sigma_2^2)$ $Y_1,Y_2\cdots Y_{n_2}$	否定域及属于哪类分布	
H_0类	μ_1、μ_2已知	$H_0:\sigma_1^2=\sigma_2^2\leftrightarrow H_1:\sigma_1^2\neq\sigma_2^2$	$\left\{\dfrac{\sum(X_i-\mu_1)^2/n_1}{\sum(Y_i-\mu_2)^2/n_2}=F_0>c_2\text{ 或 }<c_1\right\}$	$F_0\sim F(n_1,n_2)$
		$H_0:\sigma_1^2\leq\sigma_2^2\leftrightarrow H_2:\sigma_1^2>\sigma_2^2$	$\left\{\dfrac{\sum(X_i-\mu_1)^2/n_1}{\sum(Y_i-\mu_2)^2/n_2}=F_0>c\right\}$	$F_0\sim F(n_1,n_2)$
	μ_1、μ_2未知	$H_0:\sigma_1^2=\sigma_2^2\leftrightarrow H_1:\sigma_1^2\neq\sigma_2^2$	$\left\{\dfrac{\sum(X_i-\bar X)^2/(n_1-1)}{\sum(Y_i-\bar Y)^2/(n_2-1)}=F_1>c_2\text{ 或 }<c_1\right\}$	$F_1\sim F(n_1-1,n_2-1)$
		$H_0:\sigma_1^2\leq\sigma_2^2\leftrightarrow H_2:\sigma_1^2>\sigma_2^2$	$\left\{\dfrac{\sum(X_i-\bar X)^2/(n_1-1)}{\sum(Y_i-\bar Y)^2/(n_2-1)}=F_1>c\right\}$	$F_1\sim F(n_1-1,n_2-1)$
	$\sigma_1^2=\sigma_2^2$未知	$H_0:\mu_1=\mu_2\leftrightarrow H_1:\mu_1\neq\mu_2$	$\left\{\lvert T_2\rvert=\left\lvert\dfrac{\bar X-\bar Y}{\sqrt{\sum(X_i-\bar X)^2+\sum(Y_i-\bar Y)^2}}\right\rvert\sqrt{\dfrac{n_1n_2(n_1+n_2-2)}{n_1+n_2}}>c\right\}$	$T_2\sim t(n_1+n_2-2)$
		$H_0:\mu_1\leq\mu_2\leftrightarrow H_1:\mu_1>\mu_2$	$\left\{T_2=\dfrac{\bar X-\bar Y}{\sqrt{\sum(X_i-\bar X)^2+\sum(Y_i-\bar Y)^2}}\sqrt{\dfrac{n_1n_2(n_1+n_2-2)}{n_1+n_2}}>c\right\}$	$T_2\sim t(n_1+n_2-2)$
	$\sigma_1^2=\sigma_2^2$已知	$H_0:\mu_1=\mu_2\leftrightarrow H_1:\mu_1\neq\mu_2$	$\left\{\lvert M_0\rvert=\left\lvert\dfrac{\bar X-\bar Y}{\sqrt2/\sqrt n}\right\rvert>c\right\}$	$M_0\sim N(0,1)$
		$H_0:\mu_1\leq\mu_2\leftrightarrow H_1:\mu_1>\mu_2$	$\left\{M_0=\dfrac{\bar X-\bar Y}{\sqrt2/\sqrt n}>c\right\}$	$M_0\sim N(0,1)$

注：$\sigma_1^2\neq\sigma_2^2$ 无法使用广义似然比得到 UMP 检验.

自强不息 厚德载物

三、回归分析与方差分析

① 一元线性回归与正比例回归

- $y_i = a + bx_i + \varepsilon \qquad \varepsilon \in N(0, \sigma^2)$

$l_{xy} = \sum (x_i-\bar{x})(y_i-\bar{y})$，$l_{xx} = \sum (x_i-\bar{x})^2$，$l_{yy} = \sum (y_i-\bar{y})^2$，$U = \sum (\hat{y}_i-\bar{y})^2$ 回归平方和，$Q = \sum (y_i-\hat{y}_i)^2$ 残差平方和

$\hat{b} = \dfrac{l_{xy}}{l_{xx}}$，$\hat{a} = \bar{y} - \hat{b}\bar{x}$

平方和分解公式：$l_{yy} = Q + U$

用 $F = \dfrac{U}{Q/(n-2)}$ 衡量回归优秀程度，其服从 $F(1, n-2)$ 分布。

$H_0: b=0 \Leftrightarrow H_1: b \neq 0$

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}}, \quad F(r) = \frac{r^2}{1-r^2} \cdot \frac{1}{n-2}$$

r^2 用 l_{xy}, l_{xx}, l_{yy} 表示：$r^2 = \dfrac{l_{xy}^2}{l_{xx} l_{yy}} = \dfrac{U}{(l_{yy})}$

y_0 的置信区间枢轴量：$T = \dfrac{y_0 - \hat{y}_0}{\sqrt{dQ/(n-2)}} \sim (t(n-2))$，其中 $d = 1 + \dfrac{1}{n} + \dfrac{(x_0-\bar{x})^2}{l_{xx}}$

- $y = bx + e \qquad e \sim N(0, \sigma^2)$

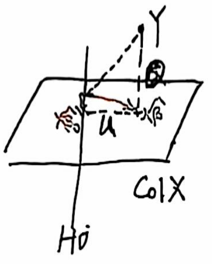
$\hat{b} = \dfrac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$ 检验 $H_0: b=0$ $\quad F = \dfrac{\hat{b}^2 \sum_{i=1}^{n} x_i^2}{Q/(n-1)} \sim F(1, n-1) \quad (Q = \sum (y_i - \hat{b}x_i)^2)$

$F > \lambda$ 时否定 H_0

② 多元线性回归

$Y = X\beta + \varepsilon \qquad \hat{\beta} = (X'X)^{-1}X'Y$，记 $\operatorname{rank} X = r$，$\operatorname{rank} H = q$

参数检验：$H_0: H\beta = 0 \leftrightarrow H_1: H\beta \neq 0$



$$\lambda = \left(\frac{\|Y - X\hat{\beta}_0\|^2}{\|Y - X\hat{\beta}\|^2}\right)^{\frac{n}{2}} = \left(1 + \frac{r-q}{n-r} F\right)^{\frac{n}{2}}$$

其中 $F \sim F(r-q, n-r)$，$\lambda > \lambda_0$ 时拒绝 H_0

四、贝叶斯估计

1、已知样本 $X_1, \cdots, X_n$，估计参数 $\theta \in \Theta$，似然函数为 L，θ 的先验分布为 $\pi(\theta)$

后验分布：

$$\pi(\theta | X_1, \cdots, X_n) = \frac{L\pi(\theta)}{\int_\Theta L\pi(\theta)\, d\theta}$$

① $X \sim B(1, p)$，$\pi(p)$：$beta(\alpha, \beta)$，$X_1, X_2 \cdots X_n$，

则 $\pi(p | X_1, X_2 \cdots X_n)$：$beta(\alpha + \Sigma X, \beta + n - \Sigma X)$

② $X \sim poisson(\lambda)$，$\pi(\lambda)$：$gamma(\alpha, \beta)$，$X_1, X_2 \cdots X_n$

则 $\pi(\lambda | X_1, X_2 \cdots X_n)$：$gamma(\alpha + \Sigma X, \beta + n)$

③ $X \sim Exp(\lambda)$，$\pi(\lambda)$：$gamma(\alpha, \beta)$，$X_1, X_2 \cdots X_n$

则 $\pi(\lambda | X_1, X_2 \cdots X_n)$：$gamma(\alpha + n, \beta + \Sigma X)$

④ $X \sim N(\mu, \sigma^2)$，σ^2 已知，μ 未知，$\pi(\mu)$：$N(\mu_0, \sigma_0^2)$，$X_1, X_2 \cdots X_n$

则 $\pi(\mu | X_1, \cdots X_n)$：$N\left(\frac{\mu_0 \sigma^2 + n\sigma_0^2 (\bar{X})^2}{\sigma^2 + n\sigma_0^2}, \frac{\sigma^2 \sigma_0^2}{\sigma^2 + n\sigma_0^2}\right)$

⑤ $X \sim N(\mu, \frac{1}{R})$，$\mu$ 已知，R 未知，$\pi(R)$：$gamma(\alpha, \beta)$，$X_1, X_2 \cdots X_n$

则 $\pi(R | X_1, \cdots, X_n)$：$gamma\left(\alpha + \frac{n}{2}, \beta + \frac{1}{2}\sum_{i=1}^{n}(X_i - \mu)^2\right)$

2、贝叶斯估计

定义损失函数 $L(\hat{\theta}, \theta)$，$\theta \in \Theta$，似然函数为 L，θ 的先验分布为 π_1，后验分布为 π_2

贝叶斯估计量是积分

$$\int_\Theta L(\hat{\theta}, \theta)\, \pi_2\, d\theta$$

取最小时的 θ 值。

自强不息 厚德载物

五、一些杂乱的定义与知识点、

1、Fischer信息量与C-R不等式

$$I(\theta) = \underbrace{E\left[\frac{\partial \ln f(x,\theta)}{\partial \theta}\right]^2}_{\text{用一阶偏导}} = \underbrace{-E\left[\frac{\partial^2 \ln f(x,\theta)}{\partial \theta^2}\right]}_{\text{用二阶偏导}}, \quad Var_\theta(\varphi(x_1 \cdots x_n)) \geq \boxed{\frac{g'(\theta)^2}{n I(\theta)}}$$

2、相合：$\lim P(|\hat\theta_n - \theta| < \varepsilon) = 1$；强相合：$P(\lim \hat\theta_n = \theta) = 1$

3、充分统计量：$L(x_1 \cdots x_n | \theta) = g[\varphi(x), \theta] h(x)$，$\varphi$为充分统计量

完全统计量：对于每个可测函数$u(\cdot)$，若 $\underline{E_\theta u(\varphi(x)) = 0}$ 可推出 $\underline{P(u(\varphi(x)) = 0) = 1}$，称$\varphi$为完全统计量

4、指数型分布

① 通式：$f(x,\theta) = S(\theta) h(x) e^{\sum C(\theta) T(x)}$

充分统计量φ是完全的的一个快速判据为 $\underline{\theta\text{有内点、}}$

② 单参数指数型分布：$f(x,\theta) = S(\theta) h(x) e^{C(\theta) T(x)}$

$H_0: \theta \leq \theta_1 \leftrightarrow H_1: \theta > \theta_1$，否定域 $\{\underline{\sum_{i=1}^n T(x) > c}\}$ 为检验水平为α的UMP

其中临界值求法：$\underline{P(\sum T(x) > c | \theta = \theta_1) = \alpha}$

5、操作特性函数 $L_W(\theta) = P(\text{接受} H_0 | \theta)$

功效函数 $P_W(\theta) = P(\text{拒绝} H_0 | \theta)$

$\}$ $L_W(\theta) + P_W(\theta) = 1$.

UMP否定域：$\forall$水平不超过α的$\tilde{W}$均有 $P_W(\theta) \geq P_{\tilde{W}}(\theta)$，称$W$为UMP否定域（$\forall \theta \in \Theta_1$

无偏性：<u>正确时被拒绝的概率小于错误时被拒绝的概率</u>

6、N-P引理.

7、minimax决策.

$$\sup_\theta R(\theta, \delta^*) \leq \sup_\theta R(\theta, \delta)$$