概率统计 (A): Final Cheatsheet

Anonymous

June 17, 2022

1 概率部分

定义 1 (二项分布). $X \sim \mathcal{B}(n,p), \mathbb{P}(X=k) = \binom{n}{k} p^k (1-p)^{n-k}, \mathbb{E}[X] = np, \mathrm{Var}[X] = np(1-p)$.

定义 2 (Poisson 分布). $X \sim \pi(\lambda), \mathbb{P}(X=k) = \mathrm{e}^{-\lambda}\frac{\lambda^k}{k!}, \mathbb{E}[X] = \lambda, \mathrm{Var}[X] = \lambda$.

引理 3. $X_n \sim \mathcal{B}(n,p_n), \lim\limits_{n\to\infty} np_n = \lambda \Rightarrow \lim\limits_{n\to\infty} \mathbb{P}(X_n = k) = \mathrm{e}^{-\lambda}\frac{\lambda^k}{k!}$.

定义 4 (负二项分布). $X \sim \mathrm{NB}(r,p), \mathbb{P}(X=k) = \binom{k+r-1}{r-1} p^r (1-p)^k, \mathbb{E}[X] = \frac{r(1-p)}{p}, \mathrm{Var}[X] = \frac{r(1-p)}{p^2}$.

定义 5 (均匀分布). $X \sim \mathcal{U}(a,b), f(x) = \frac{1}{b-a}\mathbb{1}[a \leqslant x < b], \mathbb{E}[X] = \frac{a+b}{2}, \mathrm{Var}[X] = \frac{(b-a)^2}{12}$.

定义 6 (指数分布). $X \sim \mathrm{Exp}(\lambda), f(x) = \lambda \mathrm{e}^{-\lambda x}\mathbb{1}[x \geqslant 0], \mathbb{E}[X] = \frac{1}{\lambda}, \mathrm{Var}[X] = \frac{1}{\lambda^2}$.

定义 7 (Gamma 分布). $X \sim \Gamma(\alpha,\lambda), f(x) = \frac{x^{\alpha-1}\lambda^\alpha \mathrm{e}^{-\lambda x}}{\Gamma(a)}\mathbb{1}[x \geqslant 0], \mathbb{E}[X] = \frac{\alpha}{\lambda}, \mathrm{Var}[X] = \frac{\alpha}{\beta^2}$.

定义 8 (正态分布). $X \sim \mathcal{N}(\mu,\sigma^2), f(x) = \frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}, \mathbb{E}[X] = \mu, \mathrm{Var}[X] = \sigma^2$.

命题 9 (密度变换). $f(y)\mathrm{d}y = f(x)\mathrm{d}x \Rightarrow f(y) = f(x)\cdot\frac{\mathrm{d}x}{\mathrm{d}y}$. 类似地, $f(\mathbf{y}) = f(\mathbf{x})|J(\mathbf{y})| = f(\mathbf{x})|\cdot\frac{\partial x}{\partial y}|$.

定义 10 (协方差与相关系数). $\mathrm{Cov}(X)\,Y = \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])], \rho_{XY} = \mathrm{Cov}(X)\,Y/\sqrt{\mathrm{Var}[X]\,\mathrm{Var}[Y]}$.

定义 11 (二元正态分布与多元正态分布). $X, Y \sim \mathcal{N}(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$,

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

$\mathbf{X} \sim \mathcal{N}(\mathbf{a}, B)$,

$$f(\mathbf{x}) = \frac{\mathrm{e}^{-\frac{1}{2}(\mathbf{x}-\mathbf{a})^{\mathrm{T}}B^{-1}(\mathbf{x}-\mathbf{a})}}{(2\pi)^{n/2}|B|^{1/2}}$$

定义 12 (特征函数). 随机变量 X 的特征函数为 $\psi_X(t) = \mathbb{E}\left[\mathrm{e}^{itX}\right]$.

例如, $X \sim \pi(\lambda)$ 时, $\psi_X(t) = \mathrm{e}^{\lambda(\mathrm{e}^{it}-1)}$, $X \sim \mathcal{N}(\mu,\sigma^2)$ 时, $\psi_X(t) = \mathrm{e}^{i\mu t - \sigma^2 t^2/2}$.

定理 13 (唯一性定理). 随机变量的分布函数由特征函数唯一确定.

1.1 大数定律

对于随机变量 $X_1, X_2, \cdots, X_n$, 记 $\mu_n = \frac{1}{n}\sum\limits_{i=1}^{n}\mathbb{E}[X_i]$(如果存在的话), 且满足如下条件之一:

- (Chebyshev) $\{X_i\}$ pairwise independent, 且方差有界, 即存在 C 使任意 X_i 满足 $\mathrm{Var}[X_i] \leqslant C$.
- (Markov) $\lim\limits_{n\to\infty}\frac{1}{n^2}\mathrm{Var}\left[\sum\limits_{i=1}^{n} X_i\right] = 0$.
- (Khinchin) $\{X_i\}$ i.i.d. , 期望存在.

则对于任意 $\varepsilon > 0$, 如下极限式成立

$$\lim_{n\to\infty}\mathbb{P}\left(\left|\frac{\sum_{i=1}^{n} X_i}{n} - \mu_n\right| \leqslant \varepsilon\right) = 1$$

1.2 随机变量的收敛性

定义 14 (依概率收敛). 对于一列随机变量 $X_1, X_2, \cdots, X_n, \cdots$, 如果对于 $\forall\varepsilon > 0$ 都有

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| < \varepsilon) = 1$$

则称随机变量序列 $\{X_n\}$ <u>**依概率收敛**</u>到 X, 记作 $\lim_{n\to\infty} X_n \stackrel{P}{=} X$.

记 $A_n(\varepsilon) = \{|X_n - X| \geqslant \varepsilon\}$, 则 $\{X_n\}$ 依概率收敛到 X 当且仅当 $\forall\varepsilon$, $\lim_{n\to\infty} \mathbb{P}(A_n(\varepsilon)) = 0$.

定义 15 (几乎必然收敛). 对于一列随机变量 $X_1, X_2, \cdots, X_n, \cdots$, 如果[1]

$$\mathbb{P}\left(\lim_{n\to\infty} X_n = X\right) = 1$$

则称随机变量序列 $\{X_n\}$ <u>**几乎必然收敛**</u>到 X, 记作 $\lim_{n\to\infty} X_n \stackrel{a.s.}{=} X$.

记 $A_n(\varepsilon) = \{|X_n - X| \geqslant \varepsilon\}$, 则 $\{X_n\}$ 几乎必然收敛到 X 当且仅当 $\forall\varepsilon$, $\lim_{n\to\infty} \mathbb{P}\left(\bigcup_{m\geqslant n} A_m(\varepsilon)\right) = 0$.

几乎必然收敛是比依概率收敛要严格强的性质.

定义 16 (依分布收敛). 对于一列随机变量 $X_1, X_2, \cdots, X_n, \cdots$, 如果对于 $F_X(x)$ 的每个连续点 x, 都有

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$

则称随机变量序列 $\{X_n\}$ <u>**依分布收敛**</u>到 X, 记作 $\lim_{n\to\infty} X_n \xrightarrow{d} X$. 称分布函数序列 $\{F_{X_n}(x)\}$ <u>**弱收敛**</u>到 $F_X(x)$.

依概率收敛是比依分布收敛要严格强的性质. 但依分布收敛到常数也可以推出依概率收敛.

定理 17 (连续性定理). 随机变量序列 $\{X_n\}$ 依分布收敛到 X (分布函数序列 $\{F_{X_n}(x)\}$ <u>**弱收敛**</u>到 $F_X(x)$), 当且仅当 $\{\psi_{X_n}(t)\}$ 弱收敛到 $\psi_X(t)$.

1.3 中心极限定理

定理 18 (Lindeberg-Lévy 定理). $\{X_i\}$ 独立同分布, $\mathbb{E}[X_i] = \mu, \mathrm{Var}[X_i] = \sigma^2$, 记 $\tilde{S}_n = \frac{\sum_{i=1}^n X_i - \mu}{\sqrt{n}\sigma}$, 则有 $\lim_{n\to\infty} \xrightarrow{d} Z \sim \mathcal{N}(0,1)$ (依分布收敛到标准正态分布).

利用 $\tilde{S}_n$ 的特征函数, 证明其收敛到 e^{-t^2}.

2 统计部分

定义 19 (样本均值, 方差). X 是总体, $(X_1, \cdots, X_n)$ 是取自总体的样本, 则 $\overline{X} = \sum_i X_i/n$, $S^2 = \sum_i (X_i - \overline{X})^2/(n-1)$, $\mathbb{E}\left[\overline{X}\right] = \mathbb{E}[X]$, $\mathbb{E}\left[S^2\right] = \mathrm{Var}[X]$.

定义 20 (χ^2 分布). $X_i \sim$ i.i.d. $\mathcal{N}(0,1), X = \sum_i X_i^2, X \sim \chi^2(n)$,

$$f(x) = \frac{(x/2)^{n/2-1}\mathrm{e}^{-x/2}}{2\Gamma(n/2)} \mathbb{1}[x \geqslant 0]$$

$\mathbb{E}[X] = n, \mathrm{Var}[X] = 2n$. $\mathbb{E}[1/X] = \frac{1}{n-2}$.

注 21. $\chi^2(2) = \mathrm{Exp}\left(\frac{1}{2}\right)$.

定义 22 (t 分布). $X \sim \mathcal{N}(0,1), Y \sim \chi^2(n)$, 两者独立, 则 $T = \frac{X}{\sqrt{Y/n}} \sim t(n)$,

$$f(x) = \frac{\Gamma((n+1)/2)}{\sqrt{n\pi}\Gamma(n/2)}(1 + x^2/n)^{-(n+1)/2}$$

定义 23 (F 分布). $X \sim \chi^2(n_1), Y \sim \chi^2(n_2)$, 两者独立, 则 $F = \frac{X/n_1}{Y/n_2} \sim F(n_1, n_2)$.

命题 24 (一些统计量的分布). 设总体 $X \sim \mathcal{N}(\mu_1, \sigma_1^2), Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$, 对于样本 $(X_1, \cdots, X_{n_1}), (Y_1, \cdots, Y_{n_2})$, 记 $\overline{X}, \overline{Y}, S_X^2, S_Y^2$ 分别表示两者样本均值与样本方差.

- $\overline{X} \sim \mathcal{N}(\mu_1, \sigma_1^2/n_1)$, 于是 $\frac{\overline{X}-\mu_1}{\sigma_1/\sqrt{n_1}} \sim \mathcal{N}(0,1)$. 显然.

[1] 这是啥意思?

- $(n_1-1)S_X^2/\sigma_1^2 \sim \chi^2(n_1-1)$, 且与 $\overline{X}$ 独立. 证法是构造对 $\mathbf{X}=(X_1,\cdots,X_{n_1})$ 的正交变换 A 满足

$$A=\begin{pmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & \cdots & 0 & 0 \\ \frac{1}{\sqrt{2\times3}} & \frac{1}{\sqrt{2\times3}} & -\frac{2}{\sqrt{2\times3}} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \cdots & \cdots \\ \frac{1}{\sqrt{(n-1)n}} & \frac{1}{\sqrt{(n-1)n}} & \frac{1}{\sqrt{(n-1)n}} & \cdots & \frac{1}{\sqrt{(n-1)n}} & -\frac{n-1}{\sqrt{(n-1)n}} \end{pmatrix}$$

则 $\mathbf{Y}=A\mathbf{X}$ 满足 $Y_1=\overline{X}/\sqrt{n_1}, \sum_{i=2}^{n_1} Y_i=(n_1-1)S_X^2$.

- $\frac{\overline{X}-\mu_1}{\sqrt{S_X^2/n_1}} \sim t(n-1)$. 因为 $\frac{\overline{X}-\mu_1}{\sigma_1/\sqrt{n_1}} \sim \mathcal{N}(0,1)$, 而 $(n_1-1)S_X^2/\sigma_1^2 \sim \chi^2(n-1)$, 且两者独立.

- $\frac{S_X^2/\sigma_1^2}{S_Y^2/\sigma_2^2} \sim F(n_1-1,n_2-1)$. 显然.

- 当 $\sigma_1^2=\sigma_2^2=\sigma^2$ 但未知时, 记 $S_W^2=\frac{\sum_i(X_i-\overline{X})^2+\sum_j(Y_j-\overline{Y})^2}{n_1+n_2-2}=\frac{(n_1-1)S_X^2+(n_2-1)S_Y^2}{n_1+n_2-2}$, 则 $\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{S_W^2(n_1^{-1}+n_2^{-1})}} \sim t(n_1+n_2-2)$. 注意到正态分布的线性变换仍是正态分布, 故 $\overline{X}-\overline{Y} \sim \mathcal{N}(\mu_1-\mu_2,(n_1^{-1}+n_2^{-1})\sigma^2)$, 而 $\frac{(n_1-1)S_X^2}{\sigma^2} \sim \chi^2(n_1-1)$, $\frac{(n_2-1)S_Y^2}{\sigma^2} \sim \chi^2(n_2-1)$ 说明 $\frac{(n_1+n_2-2)S_W^2}{\sigma^2} \sim \chi^2(n_1+n_2-2)$, 结合独立性即得结论.

- 当 $\sigma_1^2 \neq \sigma_2^2$ 未知时, 对于统计量 $T=\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{S_X^2}{n_1}+\frac{S_Y^2}{n_2}}}$, 当样本容量充分大时认为 T 服从标准正态分布 $\mathcal{N}(0,1)$. 当样本容量小时, 认为 T 服从 $t(k)$ 分布, 其中 $k=\min\{n_1-1,n_2-1\}$, 更精确的估计为 $k=\frac{(S_X^2/n_1+S_Y^2/n_2)^2}{(S_X^2/n_1)^2/(n_1-1)+(S_Y^2/n_2)^2/(n_2-1)}$.

2.1 参数估计

参数估计就是根据总体 X 的样本取值 $(X_1,\cdots,X_n)$ 来估计 X 的分布参数 θ.

定义 25 (矩法). 利用样本 k 阶矩 $A_k=\sum_i X_i^k/n$ 和 k 阶中心距 $B_k=\sum_i(X_i-\overline{X})^k/n$, 先写出矩关于参数的表达式, 再反求出参数关于矩的表达式.

定义 26 (极大似然法). 基于参数 θ 均匀分布的假设, $\arg\max_\theta \mathbb{P}(\theta|\mathbf{X})=\arg\max_\theta \mathbb{P}(\mathbf{X}|\theta)$, 因此考虑最大化 $L(\theta)=\mathbb{P}(x_1,x_2,\cdots,x_n$

定义 27 (无偏性, 渐进无偏性与相合性). 对参数 θ 的估计量 $\hat\theta(X_1,\cdots,X_n)$, 如果 $\mathbb{E}\left[\hat\theta\right]=\theta$, 则 $\hat\theta$ 是无偏的; 如果 $\lim_{n\to\infty}\mathbb{E}\left[\hat\theta_n\right]=0$, 则 $\hat\theta$ 是渐进无偏的; 如果 $\hat\theta_n$ 依概率收敛到 θ, 即 $\lim_{n\to\infty}\mathbb{P}\left(|\hat\theta_n-\theta|\geqslant\varepsilon\right)=0$, 则 $\hat\theta$ 是相合的.

定义 28 (有效性). 称无偏估计量 $\hat\theta_1$ 比 $\hat\theta_2$ 有效, 如果 $\mathrm{Var}\left[\hat\theta_1\right]\leqslant\mathrm{Var}\left[\hat\theta_2\right]$ (作为 $X_1,\cdots,X_n$ 的函数) 对一切 θ 成立, 且存在 θ_0 使不等号成立.

定理 29 (Cramér-Rao 不等式). 设总体 X 的概率密度函数为 $f(x;\theta)$, 参数 θ 的取值域为 $\Theta=\{\theta|a<\theta<b\}$, $u(X_1,\cdots,X_n)$ 是对 $g(\theta)$ 的一个无偏估计, 满足 **(1)** $\{x|f(x;\theta)>0\}$ 与 θ 无关, **(2)** $g'(\theta)$ 与 $\frac{\partial f(x;\theta)}{\partial\theta}$ 存在, 且对一切 $\theta\in\Theta$,

$$\frac{\partial}{\partial\theta}\int f(x;\theta)\mathrm{d}x=\int\frac{\partial}{\partial\theta}f(x;\theta)\mathrm{d}x$$
$$\frac{\partial}{\partial\theta}\int u(x_1,\cdots,x_n)\prod_i f(x_i;\theta)\mathrm{d}x_i=\int\frac{\partial}{\partial\theta}u(x_1,\cdots,x_n)\prod_i f(x_i;\theta)\mathrm{d}x_i$$

则无偏估计 u 满足

$$\mathrm{Var}\left[u\right]\geqslant\frac{[g'(\theta)]^2}{n\mathbb{E}\left[\left(\frac{\partial\ln f(x;\theta)}{\partial\theta}\right)^2\right]}$$

给出了无偏参数估计的方差下界.

定义 30 (有效估计, 渐进有效估计). 对 θ 的无偏估计 $\hat\theta$ 使 Cramér-Rao 不等式中等号成立, 则称 $\hat\theta$ 是 θ 的有效估计. 如果 $\lim_{n\to\infty}\frac{1/(nI(\theta))}{\mathrm{Var}\left[\hat\theta\right]}=1$, 则称 $\hat\theta$ 是 θ 的渐进有效估计.

定义 31 (置信区域, 置信区间). 对于待估的未知参数 θ, 设 $W(X_1,\cdots,X_2)\subseteq\Theta$ 是基于样本 $(X_1,\cdots,X_n)$ 得到的 θ 取值范围, 若满足 $\mathbb{P}(\theta\in W(X_1,\cdots,X_n))\geqslant 1-\alpha$, 则称 W 是 θ 的 $1-\alpha$ 置信区域, 其中 $1-\alpha$ 是置信度.

通常置信区域会形如一个区间, 称之为置信区间, 此时会使用区间上下界 $\hat\theta_L,\hat\theta_R$ 来刻画.

定义 32 (枢轴量). 枢轴量是关于样本与待估参数的函数, 其分布不依赖于参数. 相对应的, 统计量只是样本的函数, 其分布可以依赖参数.

可以使用枢轴量来构造置信区间. 具体的, 对于枢轴量 $G(X_1, \cdots, X_n, \theta)$, 根据其特定分布不难求出 $\mathbb{P}(a < G(X_1, \cdots, X_n, \theta) < b$ $1-\alpha$ 的区间 (a, b), 从而通过不等式变换得到 $\mathbb{P}\left(\hat{\theta}_L < \theta < \hat{\theta}_R\right) \geqslant 1-\alpha$.

正态总体均值、方差的置信区间与单侧置信限 (置信度 $1-\alpha$)

	待估参数	其他参数	W 的 分 布	置信区间	单侧置信限
一个正态总体	μ	σ^2已知	$Z=\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$	$\left(\bar{X} \pm \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\right)$	$\mu_U=\bar{X}+\frac{\sigma}{\sqrt{n}} z_\alpha$ $\mu_L=\bar{X}-\frac{\sigma}{\sqrt{n}} z_\alpha$
	μ	σ^2未知	$t=\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$	$\left(\bar{X} \pm \frac{S}{\sqrt{n}} t_{\alpha/2}(n-1)\right)$	$\mu_U=\bar{X}+\frac{S}{\sqrt{n}} t_\alpha(n-1)$ $\mu_L=\bar{X}-\frac{S}{\sqrt{n}} t_\alpha(n-1)$
	σ^2	μ未知	$\chi^2=\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$	$\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)}, \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)}\right)$	$\sigma_U^2=\frac{(n-1)S^2}{\chi^2_{1-\alpha}(n-1)}$ $\sigma_L^2=\frac{(n-1)S^2}{\chi^2_{\alpha}(n-1)}$
两个正态总体	$\mu_1-\mu_2$	σ_1^2, σ_2^2 已知	$Z=\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \sim N(0,1)$	$\left(\bar{X}-\bar{Y} \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right)$	$(\mu_1-\mu_2)_U=\bar{X}-\bar{Y}+z_\alpha\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$ $(\mu_1-\mu_2)_L=\bar{X}-\bar{Y}-z_\alpha\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$
	$\mu_1-\mu_2$	$\sigma_1^2=\sigma_2^2$ $=\sigma^2$未知	$t=\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{S_w\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t(n_1+n_2-2)$	$\left(\bar{X}-\bar{Y} \pm t_{\alpha/2}(n_1+n_2-2) S_w \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right)$	$(\mu_1-\mu_2)_U=\bar{X}-\bar{Y}+t_\alpha(n_1+n_2-2)S_w\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$ $(\mu_1-\mu_2)_L=\bar{X}-\bar{Y}-t_\alpha(n_1+n_2-2)S_w\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$
	$\frac{\sigma_1^2}{\sigma_2^2}$	μ_1, μ_2 未知	$F=\frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \sim F(n_1-1, n_2-1)$	$\left(\frac{S_1^2}{S_2^2}\frac{1}{F_{\alpha/2}(n_1-1,n_2-1)}, \frac{S_1^2}{S_2^2}\frac{1}{F_{1-\alpha/2}(n_1-1,n_2-1)}\right)$	$\left(\frac{\sigma_1^2}{\sigma_2^2}\right)_U=\frac{S_1^2}{S_2^2}\frac{1}{F_{1-\alpha}(n_1-1,n_2-1)}$ $\left(\frac{\sigma_1^2}{\sigma_2^2}\right)_L=\frac{S_1^2}{S_2^2}\frac{1}{F_{\alpha}(n_1-1,n_2-1)}$

2.2 假设检验

假设检验就是对于给出的原假设 H_0 和备择假设 H_1, 将样本取值空间划分成不交的两部分 $W, \overline{W}$, 在样本 $(x_1, \cdots, x_n) \in W$ 时拒绝原假设, 否则接受原假设. W 被称为拒绝域.

定义 33 (第 I, II 类错误). 第 I 类错误是拒绝掉真实的原假设, 第 II 类错误是接受错误的原假设. 用 α, β 分别表示两者错误率, 即

$$\alpha=\mathbb{P}(\text{reject } H_0 | H_0), \quad \beta=\mathbb{P}(\text{accept } H_0 | H_1)$$

定义 34 (Neyman-Pearson 原则, 显著性水平). 首先控制第 I 类错误的概率不超过某个常数 $\alpha \in (0,1)$, 再寻找检验, 使得第 II 类错误的概率尽可能小. 其中这里的参数 α 也被称作显著水平.

定义 35 ((样本)p-value). 样本 $(X_1, \cdots, X_n)$ 的 p-value 指的是原假设成立时, 能取到比该样本更加极端样本的概率. 当原假设是复合假设 (例如 $H_0: \mu \geqslant \mu_0$) 时, p-value 取假设集合中概率的上确界, 即

$$p=\sup_{H \in \mathcal{H}} \mathbb{P}(X' \text{ is at least as extreme as } X | H)$$

2.3 方差分析

单因素方差分析的模型为: 在 r 组不同条件下进行了总计 n 次实验, 第 j 组环境下进行了 n_j 次, 记 X_{ij} 为在第 j 组条件下进行第 i 次实验的结果, $\mu_j=\frac{1}{n_j}\sum_{i=1}^{n_j} X_{ij}$ 为第 j 组平均, $\mu=\frac{1}{n}\sum_{j=1}^{r} n_j\mu_j$ 为总平均. 记 $\delta_j=\mu_j-\mu$, 则 $\sum_{j=1}^{r} n_j\delta_j=0$,

且

$$X_{ij}=\mu+\delta_j+\varepsilon_{ij},\quad \varepsilon_{ij}\sim \text{i.i.d. } \mathcal{N}(0,\sigma^2),\sigma^2\text{未知}$$

定义 36 (偏差平方和).

$$\begin{aligned}
\text{总偏差平方和}\qquad S_T&=\sum_{j=1}^{r}\sum_{i=1}^{n_j}(X_{ij}-\mu)^2=\sum_{j=1}^{r}\sum_{i=1}^{n_j}X_{ij}^2-n\mu^2\\
\text{效应平方和}\qquad S_A&=\sum_{j=1}^{r}n_j(\mu_j-\mu)^2=\sum_{j=1}^{r}n_j\mu_j^2-n\mu^2\\
\text{误差平方和}\qquad S_E&=\sum_{j=1}^{r}\sum_{i=1}^{n_j}(X_{ij}-\mu_j)^2=\sum_{j=1}^{r}\sum_{i=1}^{n_j}X_{ij}^2-\sum_{j=1}^{r}n_j\mu_j^2
\end{aligned}$$

命题 37.

$$\begin{aligned}
S_T&=S_A+S_E\\
\mathbb{E}\left[S_T\right]&=\sum_{j=1}^{r}n_j\delta_j^2+(n-1)\sigma^2\\
\mathbb{E}\left[S_A\right]&=\sum_{j=1}^{r}n_j\delta_j^2+(r-1)\sigma^2\\
\mathbb{E}\left[S_E\right]&=(n-r)\sigma^2
\end{aligned}$$

注意到 $S_E/\sigma^2\sim\chi^2(n-r)$. 当 $\delta_1=\delta_2=\cdots=\delta_r=0$ 时, 容易发现 $S_T/\sigma^2\sim\chi^2(n-1)$ (因为 $S_T/(n-1)$ 是样本方差), 此外也可证明 $S_A/\sigma^2\sim\chi^2(r-1)$, 以及 S_A,S_E 独立, 所以 $\frac{S_A/(r-1)}{S_E/(n-r)}\sim F(r-1,n-r)$, 可用于 F 检验.

记 $S_E^j=\sum_{i=1}^{n_j}(X_{ij}-\mu_j)^2$, 则可以利用前面正交变换证明 $\overline{X}$ 与 S^2 独立的方法证明 μ_j 与 S_E^j 独立. 而 $S_E=\sum_j S_E^j$, S_A 完全由 μ_j 决定, 故两者独立.

定理 38 (Cochran). 设 $\mathbf{X}=(X_1,\cdots,X_n)^{\mathrm{T}}$, 其中 $X_1,\cdots,X_n\sim$ i.i.d. $\mathcal{N}(0,1)$. 对称矩阵 $A_1,\cdots,A_k$ 满足

$$\mathbf{X}^{\mathrm{T}}\mathbf{X}=\sum_{i=1}^{k}\mathbf{X}^{\mathrm{T}}A_i\mathbf{X}$$

记 $\operatorname{rank}(A_i)=r_i$, 则 $\mathbf{X}^{\mathrm{T}}A_i\mathbf{X}\sim\chi^2(r_i)$ 且相互独立当且仅当 $\sum_{i=1}^{k}r_i=n$.

2.4 回归分析

一元回归模型:

$$y_i=\alpha+\beta x_i+\varepsilon_i,\quad \varepsilon_i\sim\mathcal{N}(0,\sigma^2),\sigma^2\text{未知}$$

最小二乘法: 定义 $Q(\alpha,\beta)=\sum_i(y_i-\alpha-\beta x_i)^2$, 利用 $\frac{\partial Q}{\partial\alpha}=\frac{\partial Q}{\partial\beta}=0$ 得到最小二乘估计

$$\begin{aligned}
\hat\alpha&=\overline{y}-\overline{x}\hat\beta\\
\hat\beta&=s_{xy}/s_{xx}\\
s_{xy}&=\sum_i(x_i-\overline{x})(y_i-\overline{y})\\
s_{xx}&=\sum_i(x_i-\overline{x})^2
\end{aligned}$$

命题 39.

$$\begin{aligned}
\hat\beta&\sim\mathcal{N}\left(\beta,\frac{\sigma^2}{s_{xx}}\right)\\
\hat\alpha&\sim\mathcal{N}\left(\alpha,\left(\frac{1}{n}+\frac{\overline{x}^2}{s_{xx}}\right)\sigma^2\right)
\end{aligned}$$

命题 40. 记 $\hat y_i=\hat\alpha+\hat\beta x_i$, 定义残差

$$s^2=\frac{1}{n-2}\sum_i(y_i-\hat y_i)^2=\frac{s_{yy}-\hat\beta s_{xy}}{n-2}$$

则 $\mathbb{E}\left[s^2\right]=\sigma$.

类似定义三种平方和:

$$SST = \sum_i (y_i - \overline{y})^2$$
$$SSR = \sum_i (\hat{y}_i - \overline{y})^2$$
$$SSE = \sum_i (y_i - \hat{y}_i)^2$$

其中 $\frac{SSE}{\sigma^2} = \frac{(n-2)s^2}{\sigma^2} \sim \chi^2(n-2)$, 而当 $\beta = 0$ 时, 可以证明 $\frac{SSR}{\sigma^2} = \frac{s_{yy}^2}{\sigma^2 s_{xy}} \sim \chi^2(1)$ 且两者独立, 故 $\frac{SSR}{SSE/(n-2)} \sim F(1, n-2)$.