

数理统计

北京大学 龚诚欣

<https://wqgcx.github.io/>

1 绪论

- 研究有效地收集、整理和分析带有随机性的数据。
- 总体：被研究对象全体，常用随机变量 X 来表示。
- 样本：抽取的有代表性的个体。

2 估计

2.1 参数估计的方法

- 相合性：估计值依概率收敛到真实值；若几乎必然收敛则称强相合。

- 最大似然估计： $L(x_1, \dots, x_n; \theta) = \prod_{i=1}^n f(x_i, \theta)$ 。求最大值：取对数，求导。

• 矩估计： $V_k = EX^k$ ，从方程组 $g_k(\theta_1, \dots, \theta_m) = V_k$ 反解出 $\theta_k = f_k(V_1, \dots, V_m)$ ，再用样本矩估计 V_k 。

2.2 估计的优良性标准

- 无偏估计：称 f 是 $g(\theta)$ 的无偏估计，如果 $\forall \theta, E_\theta f(X_1, \dots, X_n) = g(\theta)$ 。
- 均方误差：设 f 是 $g(\theta)$ 的估计，称 $M_\theta(f) = E_\theta [f(X_1, \dots, X_n) - g(\theta)]^2$ 为 f 的均方误差。
- 若 $M_\theta(f_1) \leq M_\theta(f_2) (\forall \theta)$ ，则称 f_1 不次于 f_2 ；严格 $< (\exists \theta_0)$ 称有效。

- 定理： $S^2 := \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ 给出了 $\text{Var}(X)$ 的无偏估计。

• 充分统计量：联合密度函数可以写为 $L(x_1, \dots, x_n; \theta) = q[f(x_1, \dots, x_n), \theta] h(x_1, \dots, x_n)$ ，称 $f(x_1, \dots, x_n)$ 是 θ 的充分统计量。（等价： $P_\theta((x_1, \dots, x_n) \in A | f(x_1, \dots, x_n) = u)$ 与 θ 无关）

- 最小方差无偏估计：估计 $f(X_1, \dots, X_n)$ 无偏且在无偏估计中方差最小 ($\forall \theta$)。

• 完全性：若任何 borel 可测函数 $u(\cdot)$ ， $E_\theta u[f(X_1, \dots, X_n)] = 0 (\forall \theta)$ 就有 $P_\theta(u[f(X_1, \dots, X_n)] = 0) = 1 (\forall \theta)$ ，则称统计量 f 是完全的。

- 指数型分布： $p(x, \theta) = S(\theta) h(x) \exp\{\sum_j C_j(\theta) T_j(x)\}$ (j 往往取 1 或 2，均匀分布不是)。

• 数据预处理的优良性标准：1° 该保留的信息都保留——充分性；该丢掉的信息都丢掉——完全性。

- 若参数空间 Θ 有内点，则指数分布族中的充分统计量 $(\sum_i T_1(x_i), \dots, \sum_i T_k(x_i))$ 完全。

• BLS 定理： f 是完全的充分统计量， $h(f)$ 是 g 的无偏估计，则 $h(f)$ 是最小方差无偏估计，且在概率为 1 相等的意义下唯一。

• C-R 不等式： X 的密度函数是 $p(x, \theta)$ ， X_1, \dots, X_n 是 X 的样本， $f(X_1, \dots, X_n)$ 是 $g(\theta)$ 的无偏估计，且满足如下正则性条件：

1° $E := \{x | p(x, \theta) \neq 0\}$ 与 θ 无关；2° $g'(\theta)$ 和 $\frac{dp(x, \theta)}{d\theta}$ 都存在，且对于一切 θ 有：

$$\int_R \frac{dp(x, \theta)}{d\theta} dx = 0, \quad \int_R \dots \int_R \frac{d}{d\theta} \left[\prod_{i=1}^n p(x_i, \theta) \right] dx = 0,$$

$$\frac{d}{d\theta} \int_R \cdots \int_R f(x) \prod_{i=1}^n p(x_i, \theta) dx = \int_R \cdots \int_R f(x) \frac{d}{d\theta} \prod_{i=1}^n p(x_i, \theta) dx ;$$

$$3^\circ I(\theta) := E \left(\frac{d \ln p(x, \theta)}{d\theta} \right)^2 \text{ (fisher 信息量)}. \text{ 则有 } Var_\theta(f(X_1, \dots, X_n)) \geq \frac{[g'(\theta)]^2}{nI(\theta)}.$$

2.3 置信区间(区间估计)

• 设 $\gamma \in (0, 1)$, $f_1(X_1, \dots, X_n)$ 和 $f_2(X_1, \dots, X_n)$ 是两个统计量, $f_1 \leq f_2$ 。称 $[f_1, f_2]$ 是 $g(\theta)$ 的置信水平为 γ 的置信区间, 若对 $\forall \theta$ 均有 $P(f_1(X_1, \dots, X_n) \leq g(\theta) \leq f_2(X_1, \dots, X_n)) \geq \gamma$ 。若下确界能够取到, 则称为置信系数。

• 枢轴量方法: 寻找函数 $h(X_1, \dots, X_n, g(\theta))$ 使得这个函数的概率分布函数 $H(x)$ 与 θ 无关, 然后找 $a_1 < a_2$ 使得 $H(a_2) - H(a_1) \geq \gamma$, 再解不等式 $a_1 \leq h \leq a_2$, 得到 f_1 和 f_2 。这里的函数 h 称为枢轴量。

• n 个自由度的 χ^2 分布: $p_n(x) = I_{\{x>0\}} \frac{1}{2^{\frac{n}{2}} \Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} e^{-\frac{1}{2}x}$ 。

• $X_1, \dots, X_n \text{ i.i.d. } \sim N(0, 1)$, 则 $\xi = X_1^2 + \dots + X_n^2 \sim \chi^2(n)$, $E\xi = n$, $D\xi = 2n$ 。

• 定理: $X_1, \dots, X_n \text{ i.i.d. } \sim N(\mu, \sigma^2)$, 则

$$1^\circ \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right); \quad 2^\circ \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2 \sim \chi^2(n-1);$$

$$3^\circ \bar{X} \text{ 和 } \sum_{i=1}^n (X_i - \bar{X})^2 \text{ 相互独立}.$$

• n 个自由度的 t 分布: $p_n(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right) \sqrt{n\pi}} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}$ 。

• $X \sim N(0, 1), Y \sim \chi^2(n)$, X, Y 独立, 则 $\frac{X}{\sqrt{\frac{1}{n}Y}} \sim t(n)$ 。

• 例: 正态分布已知方差估计均值: $\frac{1}{\sigma} \sqrt{n}(\bar{X} - \mu) \sim N(0, 1)$;

估计方差: $\frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2 \sim \chi^2(n-1)$;

未知方差估计均值: $T = \frac{\sqrt{n}(\bar{X} - \mu)}{S} \sim t(n-1)$, $S = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}$ 。

• 统计量方法: 设 $f(X_1, \dots, X_n)$ 是广义实值统计量, 令 $G(c, \theta) = P_\theta(f(X_1, \dots, X_n) \geq c)$, $H(c, \theta) = P_\theta(f(X_1, \dots, X_n) > c)$ 。给定 $0 < \gamma < 1$, 令 $g_L(c) = \inf\{g(\theta) | G(c, \theta) > 1 - \gamma\}$, $g_U(c) = \sup\{g(\theta) | H(c, \theta) < \gamma\}$, 则:

- 1° $g_L(f(X_1, \dots, X_n))$ 是 $g(\theta)$ 置信水平为 γ 的置信下限, $P_\theta(g(\theta) \geq g_L(f(X_1, \dots, X_n))) \geq \gamma$;
 2° $g_U(f(X_1, \dots, X_n))$ 是 $g(\theta)$ 置信水平为 γ 的置信上限, $P_\theta(g(\theta) \leq g_U(f(X_1, \dots, X_n))) \geq \gamma$;
 3° $[g_L(f(X_1, \dots, X_n)), g_U(f(X_1, \dots, X_n))]$ 或 $[g_U, g_L]$ 给出了置信水平为 $2\gamma-1$ 的置信区间。

2.4 分布函数与密度函数的估计

• 经验分布函数: $F_n(x) = \frac{1}{n} \sum_{i=1}^n I_{(-\infty, x]}(X_i)$ 。

• G-C 定理: 设 $D_n = \sup_x |F_n(x) - F(x)|$, 则 $P(\lim_n D_n = 0) = 1$ 。

• 直方图法: $R_n(a, b)$ 表示落在区间 $(a, b]$ 的个数, 积分 $\int_a^b p(x) dx$ 可以用频率来估计。

由微分中值定理, 可以用 $\frac{R_n(a, b)}{n(b-a)}$ 作为 $p(x_0)$ 的估计值。

• 核估计法: 设 $K(x) \geq 0$, $\int K(x) dx = 1$, 称 $f_n(x) = \frac{1}{nh} \sum_{i=1}^n K(\frac{x-x_i}{h})$ 为 $f(x)$ 的核估计,

$h > 0$ 称为窗宽。窗宽 h 越小, 说明越重视靠近 x 的数据。数据量大, h 可减小。样本固定时, 窗宽 h 大, 核估计平滑; 窗宽 h 小, 核估计波动大。

3 假设检验

3.1 问题的提法

- 第一类错误: 以真为假; 第二类错误: 以假为真。
- $L_w(\theta) = P(\text{接受 } H_0 | \theta) = P((x_1, \dots, x_n) \notin W | \theta)$, $\rho_w(\theta) = P(\text{拒绝 } H_0 | \theta) = P((x_1, \dots, x_n) \in W | \theta)$ 。前者称为 W 的操作特性函数, 后者称为功效函数。
- 称 $\sup_{\theta \in \Theta_0} \rho_w(\theta)$ 为 W 的检验水平。

• $H_0: \theta \in \Theta_0 \longleftrightarrow H_1: \theta \in \Theta_1$, 通常情况 $\Theta_1 = \Theta - \Theta_0$ 。【往往取 H_0 是闭集】

• 无偏否定域: 若 W 的水平为 α , 且对于一切 $\theta \in \Theta_1$, 都有 $\rho_w(\theta) \geq \alpha$, 则称 W 是检验水平为 α 的无偏否定域。

• UMP 否定域: 若 W 是水平为 α 的否定域且对一切水平不超过 α 的否定域 W' 成立 $\rho_w(\theta) \geq \rho_{w'}(\theta) (\theta \in \Theta_1)$, 则称 W 为一致最大功效否定域。【没有无偏要求】

• UMPU (一致最大功效无偏) 否定域: W 是水平为 α 的无偏否定域, 且对于任何水平为 α 的无偏否定域 W' 都有 $\rho_w(\theta) \geq \rho_{w'}(\theta) (\theta \in \Theta_1)$ 。

3.2 N-P 引理及似然比检验法

• 考虑检验问题 $H_0: \theta = \theta_1 \longleftrightarrow H_1: \theta = \theta_2$, 下记 $x = (x_1, \dots, x_n)$ 。

• N-P 引理: 给定 $\alpha \in (0, 1)$, 设 $W_0 = \{x: L(x, \theta_2) > \lambda_0 L(x, \theta_1)\}$ 适合 $\int_{W_0} L(x, \theta_1) = \alpha$, 则 W_0 是一致最大功效否定域。

• 设 X 的分布密度函数是 $p(x, \theta_i)$, X 的可能值集合 $\{x: p(x, \theta_i) > 0\}$ 与 i 无关, $\lambda(x) = L(x, \theta_2)/L(x, \theta_1)$, 设 $X = (X_1, \dots, X_n)$ 是样本, 若 $\lambda(X)$ 在 θ_1 下的分布函数是连续的, 则对于任意 $\alpha \in (0, 1)$, 存在 $\lambda_0 > 0$, 使得 $W_0 = \{x | \lambda(x) > \lambda_0\}$ 是水平为 α 的唯一最大功效 UMP 否定域。这里的唯一是指相差一个 lebesgue 零测集。

• 无偏: 上述定理中的 $\rho_{W_0}(\theta_2) \geq \rho_{W_0}(\theta_1)$ 。任何 UMP 检验都是 UMPU 检验。

3.3 单参数情形的假设检验

• X 的可能值集合是 \mathcal{X} , 称 X 服从单参数指数型分布, 若 X 的分布函数 $p(x, \theta)$ 有下列表达式: $p(x, \theta) = S(\theta)h(x)e^{Q(\theta)V(x)}$, 其中 $\theta \in (a, b)$, $h(x), S(\theta) > 0$, $Q(\theta)$ 单调增。

• 考虑检验问题 $H_0: \theta \leq \theta_1 \longleftrightarrow H_1: \theta > \theta_1$, 对于 $a \in (0, 1)$, 若存在 C 满足 $P(\sum_i V(X_i) > C | \theta_1) = a$, 则 $W_0 = \{(x_1, \dots, x_n) | \sum_i V(X_i) > C\}$ 是检验水平为 a 的一致最大功效否定域。

• 设 X 有分布密度 $p(x, \theta) = S(\theta)h(x)e^{Q(\theta)V(x)}$, $S(\theta) > 0$, $h(x) \geq 0$, $Q(\theta)$ 严格单调增, 考虑检验问题 $H_0: \theta \notin (\theta_1, \theta_2) \longleftrightarrow H_1: \theta \in (\theta_1, \theta_2)$, $W_0 = \{C_1 < \sum_i V(X_i) < C_2\}$, 若 $P(X \in W_0 | \theta_1) = P(X \in W_0 | \theta_2) = a$, 则 W_0 是水平为 a 的一致最大功效否定域。

• 设 X 有分布密度 $p(x, \theta) = S(\theta)h(x)e^{Q(\theta)V(x)}$, $S(\theta) > 0$, $h(x) \geq 0$, $Q(\theta)$ 严格单调增连续, 考虑检验问题 $H_0: \theta \in [\theta_1, \theta_2] \longleftrightarrow H_1: \theta \notin [\theta_1, \theta_2]$, $W_0 = \{\sum_i V(X_i) < C_1 \text{ 或 } > C_2\}$, 若 $C_1 < C_2$ 使得 $P(X \in W_0 | \theta_1) = P(X \in W_0 | \theta_2) = a$, 则 W_0 是水平为 a 的一致最大功效无偏否定域 (此时 UMP 否定域不存在)。

• 设 X 有分布密度 $p(x, \theta) = S(\theta)h(x)e^{Q(\theta)V(x)}$, $S(\theta) > 0$, $h(x) \geq 0$, $Q'(\theta) > 0$, 考虑检验问题 $H_0: \theta = \theta_0 \longleftrightarrow H_1: \theta \neq \theta_0$, $W_0 = \{\sum V(X_i) < C_1 \text{ 或 } > C_2\}$, 若 $C_1 < C_2$ 使得 $P(X \in W_0 | \theta_0) = a$, $E_{\theta_0}(I_{W_0}(X_1, \dots, X_n) \sum_i V(X_i)) = a E_{\theta_0} \sum_i V(X_i)$, 则 W_0 是水平为 a 的一致最大功效无偏否定域 (此时 UMP 否定域不存在)。

3.4 广义似然比检验

• 考虑检验问题 $H_0: \theta \in \Theta_0 \longleftrightarrow H_1: \theta \in \Theta - \Theta_0$, 令 $L(\Theta) = \sup_{\theta \in \Theta} L(x, \theta)$ 。

• 定义 $\lambda(x) = L(\Theta)/L(\Theta_0)$ 为样本值 x 的广义似然比, 取否定域 $W_0 = \{x: \lambda(x) > \lambda_0\}$, 满足 $\sup P(x \in W_0 | \theta) (\theta \in \Theta_0) = a$, 这里 a 是预先给定的检验水平。 $\lambda(x)$ 是充分统计量的函数, 因此 $W_0 = \{x: f(x) \in B\}$ 。下面讨论正态分布。

• 检验问题 $H_0: \mu = \mu_0 \longleftrightarrow H_1: \mu \neq \mu_0$, 方差已知, 广义似然比 $\lambda(x) = e^{\frac{n}{2\sigma^2}(\bar{x} - \mu_0)^2}$, 否定域 $W = \{x: |\bar{x} - \mu_0| > C\}$ 。

• 检验问题 $H_0: \mu = \mu_0 \longleftrightarrow H_1: \mu \neq \mu_0$, 方差未知, 广义似然比 $\lambda(x) = \left(1 + \frac{T^2}{n-1}\right)^{-\frac{n}{2}}$, 其

中 $T = \frac{\sqrt{n(n-1)}(\bar{x} - \mu_0)}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} \sim t(n-1)$, 否定域 $W = \{x: |T| > C\}$ 。

• 检验问题 $H_0: \mu \leq \mu_0 \longleftrightarrow H_1: \mu > \mu_0$, 方差未知, 否定域 $W = \{x: T > C_1\}$ 。

• 检验问题 $H_0: \mu \geq \mu_0 \longleftrightarrow H_1: \mu < \mu_0$, 方差未知, 否定域 $W = \{x: T < C_2\}$ 。

• 检验问题 $H_0: \sigma^2 = \sigma_0^2 \longleftrightarrow H_1: \sigma^2 \neq \sigma_0^2$, 均值未知, 否定域 $W = \{x: u > C_2 \text{ 或 } < C_1\}$,

其中 $u = \frac{1}{\sigma_0^2} \sum_{i=1}^n (x_i - \bar{x})^2$

• 检验问题 $H_0: \sigma^2 \leq \sigma_0^2 \longleftrightarrow H_1: \sigma^2 > \sigma_0^2$, 均值未知, 否定域 $W = \{x: u > C_2\}$ 。

• 检验问题 $H_0: \sigma_1^2 = \sigma_2^2 \longleftrightarrow H_1: \sigma_1^2 \neq \sigma_2^2$ (两个正态分布样本), 否定域 $W = \{(x, y): F <$

$C_1 \text{ 或 } > C_2\}$, 其中 $F = \frac{\sum_{i=1}^{n_1} (x_i - \bar{x})^2 / (n_1 - 1)}{\sum_{i=1}^{n_2} (y_i - \bar{y})^2 / (n_2 - 1)}$ 。

- 设 $X \sim \chi^2(n_1), Y \sim \chi^2(n_2), X$ 和 Y 独立, $Z = \frac{X/n_1}{Y/n_2}$ 服从第一自由度 n_1 , 第二自由度

n_2 的 F 分布 $F(n_1, n_2)$, 密度函数 $f_{n_1, n_2}(u) = I_{\{u>0\}} \frac{\Gamma(\frac{n_1+n_2}{2})}{\Gamma(\frac{n_1}{2})\Gamma(\frac{n_2}{2})} \left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}} u^{\frac{n_1}{2}-1} \left(1 + \frac{n_1}{n_2}u\right)^{-\frac{n_1+n_2}{2}}$

- $X \sim F(n_1, n_2)$, 则 $1/X \sim F(n_2, n_1)$; $T \sim t(n)$, 则 $T^2 \sim F(1, n)$ 。

- 记 $T = \frac{\bar{x} - \bar{y}}{\sqrt{\sum_{i=1}^{n_1} (x_i - \bar{x})^2 + \sum_{i=1}^{n_2} (y_i - \bar{y})^2}} \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}}$ 。若 $X \sim N(\mu_1, \sigma_1^2), Y \sim N(\mu_2, \sigma_2^2)$,

X 和 Y 独立, $\mu_1 = \mu_2, \sigma_1^2 = \sigma_2^2$, 则 $T \sim t(n_1 + n_2 - 2)$ 。

- 假设检验与置信区间的关系: 若由数据, 水平 α 下不否定 $H_0: \theta = \theta_0$, 则充分必要地, 相应的 $1-\alpha$ 水平的置信区间包含 θ_0 。

3.5 临界值与 p 值

- 当 H_0 成立时, 产生如观测数据同样奇怪或更奇怪的数据的概率。p 值在适中范围内表示正常, 太小表示 H_0 应受强烈怀疑, 但不简单回答否定或不否定。

3.7 拟合优度检验

- χ^2 检验法: 检验问题 $H_0: F(x) = F_0(x) \leftrightarrow H_1: F(x) \neq F_0(x)$ 。设 X_1, \dots, X_n 是来自 X 的样本, 在实轴上取 m 个点 $t_1 < \dots < t_m$ 将实轴分为 $m+1$ 段, v_i 表示落入第 i 段的

个数, v_i/n 表示频率, p_i 表示相应概率, 统计量 $V = \sum_{i=1}^{m+1} \frac{(v_i - np_i)^2}{np_i}$ 服从 m 个自由

度的 χ^2 分布的密度函数。可以找到 c 使得 $\int_c^{+\infty} g_m(y) dy = \alpha$, 于是 H_0 的否定域是

$W_0 = \{V > c\}$ 。【本质是: 验证在 $t_i < t_{i+1}$ 上的概率是不是 p_i 】

- 检验问题 $H_0: F(x) \in \{F_0(x, \theta_1, \dots, \theta_k): (\theta_1, \dots, \theta_k) \in \Theta\}$, 利用多项分布求出 $\theta_1, \dots, \theta_k$ 的

最大似然估计 $\theta_{10}, \dots, \theta_{k0}$, 即 $\sum_{i=1}^{m+1} \frac{v_i}{p_i(\theta_1, \dots, \theta_k)} \frac{\partial p_i(\theta_1, \dots, \theta_k)}{\partial \theta_j} = 0$ 。设 $p_i = p_i(\theta_{10}, \dots, \theta_{k0})$,

可以证明 $V = \sum_{i=1}^{m+1} \frac{(v_i - np_i)^2}{np_i}$ 服从 $m-k$ 个自由度的 χ^2 分布, k 是未知参数的个数。

实际应用中, 求解 $\theta_{10}, \dots, \theta_{k0}$ 太麻烦, 采用下述做法: 先利用最大似然估计找出参数估计值, 得到分布函数, 再利用基本的 χ^2 (注意是 $m-k$ 个自由度) 检验法。

- 列联表的独立性检验: X 的可能取值是 $1 \sim s$, Y 的可能取值是 $1 \sim t$, “ X 取 i, Y

取 j ” 发生了 n_{ij} 次, 记 $n_{i.} = \sum_{j=1}^t n_{ij}, n_{.j} = \sum_{i=1}^s n_{ij}$, 待检验的假设 $H_0: p_{ij} = p_i q_j$, 首先在

H_0 成立的情况下寻找最大似然估计, $\ln L = \sum_{i=1}^s \sum_{j=1}^t n_{ij} \ln p_i + n_{ij} \ln q_j$, 从而 $p_i = \frac{n_{i.}}{n}$,

$q_j = \frac{n_{.j}}{n}$, 研究统计量 $V = \sum_{i,j} \frac{(n_{ij} - np_i q_j)^2}{np_i q_j}$, V 的极限分布是 $\chi^2((s-1)(t-1))$, 可以找

到 $P(V > c) = \alpha$, 因此可以取否定域 $\{W: V > c\}$ 。 $V = n \left(\sum_{i=1}^s \sum_{j=1}^t \frac{n_{ij}^2}{n_i n_{.j}} - 1 \right)$ 。

• Kolmogorov 检验: 检验问题 $H_0: F(x) = F_0(x)$, 先求出经验分布函数 $F_n(x)$, 计算 $D_n = \sup_{x \in R} |F_n(x) - F_0(x)|$, 若 H_0 成立, 则 $P(\lim_n D_n = 0) = 1$ (经验分布函数 $F_n(x)$ 一致收敛真实分布函数), 取否定域 $W_0 = \{D_n > c\}$ 。可以证明, 若 X 的分布函数连续,

则 $\lim_{n \rightarrow \infty} P(\sqrt{n} D_n \leq x) = Q(x) := I_{\{x > 0\}} \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2 x^2}$ 。

4 回归分析与线性模型

4.1 引言

- 变量之间的关系: 确定性关系、相关关系。
- 回归分析: 预测和控制。

4.2 一元线性回归

• 最小二乘法: $\hat{a}, \hat{b} = \arg \min_{a,b} Q(a,b) = \sum_{i=1}^n \{y_i - (a + bx_i)\}^2$, $\hat{b} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$,

$\hat{a} = \bar{y} - \hat{b}\bar{x}$ 。其中 $l_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2$ 是总离差平方和, $Q = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ 是残差平方

和, $U = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ 是回归平方和。最小二乘估计下, $l_{yy} = Q + U$ 。用比值 U/Q

来衡量线性关系的可信程度, 比值越大可信度越高。

• 设数据 (x_i, y_i) 有结构 $y_i = a + bx_i + e_i$, 检验假设 $H_0: b = 0$, 若否定 H_0 , 则说明 y 与 x 之间有线性关系, 叫做相关性检验。当 H_0 被否定时, 回归方程称为显著的。在

H_0 和 $e_i \sim N(0, \sigma^2)$ 条件下, $F = \frac{U}{Q/(n-2)}$ 服从自由度 $(1, n-2)$ 的 F 分布, 取否定域

$W = \{F > c\}$ 即可, c 是 $F(1, n-2)$ 的 $1-\alpha$ 分位数。计算 F 时, 可以用 $U = \hat{b}l_{xy}$, $Q = l_{yy} - U$ 。

• \hat{a}, \hat{b} 是 a, b 的无偏估计 (任意分布)。当假设 $e_i \sim N(0, \sigma^2)$ 成立时, σ^2 的 MLE 是 $Q(\hat{a}, \hat{b})/n$, 无偏估计是 $Q(\hat{a}, \hat{b})/(n-2)$ 。

• 相关系数: $r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$ (把 y 和 x 看成随机变量, 其相关系数的矩估计)。

对于一元线性回归, 还有 $r^2 = \frac{U}{l_{yy}} = 1 - \frac{Q}{l_{yy}}$, $r = \hat{b} \sqrt{\frac{l_{xx}}{l_{yy}}}$ 。

• 预测: 在正态性假设下, 随机变量 $T = \frac{y_0 - \hat{y}_0}{\sqrt{dQ/(n-2)}}$ 服从 n-2 个自由度的 t 分布,

$d = 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{l_{xx}}$, $P(|T| \leq c) = 1 - \alpha$, y_0 的 1- α 水平置信区间是 $[\hat{y}_0 - c\sqrt{dQ/(n-2)}, \hat{y}_0 + c\sqrt{dQ/(n-2)}]$ 。当 x_0 变化时, 上下限轨迹构成双曲线。

• 控制: 解不等式 $\hat{y}_0 - c\sqrt{dQ/(n-2)} \geq A, \hat{y}_0 + c\sqrt{dQ/(n-2)} \leq B$ 得到 $x_0 \in [c_1, c_2]$ 。

• 一元齐次线性回归: 考虑 $Y = \mathbf{b}x + e$, 随机项 e 满足 $Ee = 0$, $\hat{b} = \sum_{i=1}^n x_i y_i / \sum_{i=1}^n x_i^2$ 叫

作最小二乘估计。检验假设 $H_0: \mathbf{b} = 0$, 统计量 $F = \frac{(\hat{b})^2 \sum_{i=1}^n x_i^2}{Q/(n-1)}$ 服从 (1, n-1) 的 F 分布,

因此有临界值 $P(F > c) = \alpha$, 取否定域 $W_0 = \{F > c\}$ 即可。

4.3 线性模型的参数估计

• 多元线性回归: $y = \sum_{i=1}^p \beta_i x_i + e$, 常假定: A: $Ee_i = 0, Ee_i e_j = 0, Ee_i^2 = \sigma^2$, B: e_1, \dots

i.i.d. 且 $e_i \sim N(0, \sigma^2)$ 。用矩阵表达为 $Y = X\beta + e$, X 是 $n \times p$ 矩阵, $n > p$, A 表达为 $Ee = 0$, $\text{Cov}(Y, Y) = \sigma^2 I$, B 表达为 $e \sim N(0, \sigma^2 I)$ 。

• 最小二乘估计: $\hat{\beta} = \arg \min_{\beta} Q(\beta) = \|Y - X\beta\|^2 \Leftrightarrow X^T X \hat{\beta} = X^T Y$ 。

• 定理: $\text{rank}(X) = p$, 假设 A 成立, 则 1° $E\hat{\beta} = \beta$; 2° $\text{Cov}(\hat{\beta}, \hat{\beta}) = \sigma^2 (X^T X)^{-1}$; 3°

$EQ(\hat{\beta}) = (n-p)\sigma^2$ 。这说明 $\hat{\beta}$ 无偏, $\frac{Q(\hat{\beta})}{n-p}$ 是 σ^2 的无偏估计。若 X 不满秩, 未必有

无偏估计。

• 线性可估性: $c^T \beta$ 是线性可估的, 若存在 Y 的线性函数 $a^T Y$ 使得 $Ea^T Y = c^T \beta$ 。

• 假定 A 成立, $c^T \beta$ 线性可估当且仅当 c^T 是 X 的行的线性组合。

• 高斯-马尔可夫: 对于线性模型 $Y = X\beta + e$, 假定 A 成立, $\hat{\beta}$ 是 β 的最小二乘估计,

若 $c^T\beta$ 线性可估, 则 $c^T\hat{\beta}=(a^*)^TY, a^*\in\mu(X)$ 必为 $c^T\beta$ 的唯一最小方差线性无偏估计。

- 若 $\hat{\beta}$ 是 β 的最小二乘估计, 则称 $c^T\hat{\beta}$ 是 $c^T\beta$ 的最小二乘估计。

- 对于线性模型 $Y=X\beta+e$, 假定 A 成立, $\text{rank}(X)=r$, 则 $\frac{Q(\hat{\beta})}{n-r}$ 是 σ^2 的无偏估计。

- 带约束的线性模型: $Y=X\beta+e, H\beta=r_0$ 。消去多余参数法: 解方程 $H\beta=r_0$, 将所有参数用无约束的参数表示出来, 再利用最小二乘法; 拉格朗日乘子法: $\hat{\beta}$ 是最小二乘估计 \Leftrightarrow 存在 $s\times 1$ 向量 c 使得 $X^TX\hat{\beta}-H^Tc=X^TY$ 。这些方法估计的 θ_i 比用 y_i 直接估计方差更小, 故称平滑。

- 进一步讨论: $|X^TX|\rightarrow 0$, 估计不稳定; 可能受个别数据较大影响。

4.4 线性模型的假设检验

- 给定线性模型 $Y=X\beta+e$, X 是已知 $n\times p$ 矩阵, β 未知 p 维向量, $e\sim N(0, \sigma^2 I)$, Y 是观测项, $\text{rank}(X)=r$ 。考虑检验问题 $H_0: H\beta=0$, H 是 $s\times p$ 矩阵。令 $W=\mu(X)$, $W_0=\mu(X)|H\beta=0$, $\dim W_0=q<r$, 则 H_0 当且仅当 $\xi:=EY\in W_0$ 。 $\hat{\xi}_0, \hat{\xi}$ 是 Y 在 W_0 和 W

上的投影, 广义似然比 $\lambda = \frac{\|Y - \hat{\xi}_0\|^n}{\|Y - \hat{\xi}\|^n} = \left(1 + \frac{\|\hat{\xi} - \hat{\xi}_0\|^2}{\|Y - \hat{\xi}\|^2}\right)^{\frac{n}{2}}$, 因此否定域 $W_0 = \{\lambda > \lambda_0\}$

$\Leftrightarrow \frac{\|\hat{\xi} - \hat{\xi}_0\|^2}{\|Y - \hat{\xi}\|^2} > \lambda_1$ 。令 $F = \frac{\|\hat{\xi} - \hat{\xi}_0\|^2 / (r-q)}{\|Y - \hat{\xi}\|^2 / (n-r)}$, 否定域 $\{F > c\}$ 。在 H_0 成立时, F 的分布是 $F(r-q, n-r)$ 。

- 称 $\hat{e} = Y - X\hat{\beta}$ 为残差, $Q = \|\hat{e}\|^2$ 为残差平方和。对于线性模型 $Y=X\beta+e$, 假定 B 成立, $\hat{\beta}$ 是 β 的最小二乘估计, 则 $X\hat{\beta}$ 和残差 \hat{e} , 残差平方和 Q 独立, $Q/\sigma^2 \sim \chi^2(n-r)$ 。

- 对于线性模型 $Y=X\beta+e$, 假定 B 成立, $c^T\beta$ 是 β 的可估线性组合, $\hat{\beta}$ 是 β 的最小二乘估计, Q 是残差平方和, 则 $c^T\hat{\beta}$ 和 Q 相互独立; 若 X 满秩, 则 $\hat{\beta}$ 与 Q 独立, 且 $\hat{\beta} \sim N(\beta, \sigma^2(X^TX)^{-1})$ 。

- 设 a^TY 是 $c^T\beta$ 的无偏估计且 $a\in\mu(X)$, 则称 a 是 c 的伴随元。

- 对于线性模型 $Y=X\beta+e$, 假定 B 成立, 并设 $c^T\beta$ 可估, 则 $\frac{c^T(\hat{\beta}-\beta)}{\hat{\sigma}\|a\|} \sim t(n-r)$,

其中 a 是 c 的伴随元, r 是 X 的秩, $\hat{\sigma} = \sqrt{\frac{Q}{n-r}}$ 。

• 若 $c^T\beta$ 可估, 要检验 $H_0: c^T\beta=r_0$, 用统计量 $t = \frac{c^T\hat{\beta}-r_0}{\hat{\sigma}\|a\|} \sim t(n-r)$ (在 H_0 条件下),

否定域为 $\{|t|>c\}$ 。要求 $c^T\beta$ 的置信区间, 由于 $\frac{c^T(\hat{\beta}-\beta)}{\hat{\sigma}\|a\|} \sim t(n-r)$, $P(|t|>c)=\alpha$,

1- α 置信水平区间是 $[c^T\hat{\beta}-c\|a\|\hat{\sigma}, c^T\hat{\beta}+c\|a\|\hat{\sigma}]$ 。

• $c^T\beta$ 可估, $Y_0=c^T\beta+e_0$, 则 $T = \frac{c^T\hat{\beta}-Y_0}{\hat{\sigma}\sqrt{\|a\|^2+1}} \sim t(n-r)$ 。【预测 Y_0 】

4.5 回归分析 (要求数据矩阵满秩)

• 假设检验: $H_0: H\beta=0$ 。 $\|X\hat{\beta}-X\hat{\beta}_0\| = \hat{\beta}^T H^T [H(X^T X)^{-1} H^T]^{-1} H\hat{\beta}$, 其中 $\hat{\beta}$ 是 LSE, $\hat{\beta}_0$ 是 H_0 条件下的 LSE。

• 当线性检验通过后, 才能检验回归系数是否为 0。X 有 h 个取值, $H_0: EY_i=X_i\beta$,

$X=X_i$ 时有 n_i 次观测 Y_{ij} , 则 $Q = \sum_{i=1}^h \sum_{j=1}^{n_i} (Y_{ij} - \mu_i)^2 = \sum_{i=1}^h \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 + \sum_{i=1}^h (\bar{Y}_i - \mu_i)^2 =$

Q_1+Q_2 , Q_1 是随机误差, Q_2 是偏离线性模型的刻画。

• 当 H_0 成立时, $F = \frac{Q_2/(h-p)}{Q_1/(n-h)} \sim F(h-p, n-h)$ 。

• 残差分析: 去判别假定 B 是否成立。 e 是残差, 在假定 B 下服从正态分布, 则

$\text{Cov}(e, e) = \sigma^2(I-P)$, $Ee=0$, 其中 $P=X(X^T X)^{-1}X^T$ 。令 $\hat{\sigma} = \sqrt{\frac{Q}{n-p-1}} = \sqrt{\frac{1}{n-p-1} \sum_{i=1}^n (\hat{e}_i)^2}$

$\gamma_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1-p_{ii}}}$, 后者叫作学生化残差, 只要 n 相当大, γ_i 近似独立同分布服从标

准正态。这说明大致有 $[0.95n]$ 个 $|\gamma_i| \leq 2$, 如不满足, 应拒绝假设 B; 满足则一般应接受假设 B。

5 试验设计与方差分析

5.1 全面试验的方差分析

• 仅有一个因素 A, 可取 s 个水平 A_1, \dots, A_s , 目标判断因素 A 对指标 Y 是否有影响, 如果有哪个水平更好。对每个水平平均安排 r 次试验, 第 i 个水平的第 j 个结果为 $Y_{ij}=\mu_i+e_{ij}$, $\{e_{ij}\}$ 独立同分布 $\sim N(0, \sigma^2)$, 假设检验 $H_0: \mu_1=\dots=\mu_s$ 。

• 当 $s \geq 3$ 时, 记 \bar{Y}_i 为水平 i 下 Y 的均值, \bar{Y} 是总平均。当 H_0 成立时, 应相差不

大。总变差分解: $\sum_{i=1}^s \sum_{j=1}^r (Y_{ij} - \bar{Y})^2 = \sum_{i=1}^s \sum_{j=1}^r (Y_{ij} - \bar{Y}_i)^2 + r \sum_{i=1}^s (\bar{Y}_i - \bar{Y})^2 = S_C + S_A$ 。 S_C 刻画

了随机误差方差 σ^2 的大小， S_A 刻画了因素 A 对 Y 的影响。 H_0 成立时 S_A 应当比较小。取统计量 $F = \frac{S_A/(s-1)}{S_c/s(r-1)} \sim F(s-1, s(r-1))$ 。

• 两因素试验的方差分析：A 有 s 个水平，B 有 t 个水平，每一组安排 r 次试验，数据为 Y_{ij1}, \dots, Y_{ijr} ，模型为 $Y_{ijk} = \mu_{ij} + \varepsilon_{ijk}$ ， ε_{ijk} 独立同分布 $\sim N(0, \sigma^2)$ 。假设检验有：A 对 Y 有无影响，B 对 A 有无影响，是否存在 A 和 B 的交互作用。参数变换

$$\mu = \frac{1}{st} \sum_{i=1}^s \sum_{j=1}^t \mu_{ij}, \alpha_i = \frac{1}{t} \sum_{j=1}^t \mu_{ij} - \mu, \beta_j = \frac{1}{s} \sum_{i=1}^s \mu_{ij} - \mu, \lambda_{ij} = \mu_{ij} - \alpha_i - \beta_j - \mu. \alpha_i \text{ 称为 A 的}$$

主效应， β_j 为因素 B 的主效应， λ_{ij} 称为 A 和 B 的交互作用。模型可以表示为 $Y_{ijk} = \mu + \alpha_i + \beta_j + \lambda_{ij} + \varepsilon_{ijk}$ 。 $H_1: \alpha_1 = \dots = \alpha_s = 0$ ， $H_2: \beta_1 = \dots = \beta_t = 0$ ， $H_3: \lambda_{11} = \dots = \lambda_{st} = 0$ 。

$$S_r = \sum_{i,j,k} (Y_{ijk} - \bar{Y})^2 = \sum_{i,j,k} (Y_{ijk} - \bar{Y}_{ij})^2 + r \sum_{i,j} (\bar{Y}_{ij} - \bar{Y}_i - \bar{Y}_j + \bar{Y})^2 + tr \sum_i (\bar{Y}_i - \bar{Y})^2 + sr \sum_j (\bar{Y}_j - \bar{Y})^2$$

$$= S_C + S_{A \times B} + S_A + S_B. H_1 \text{ 成立时, } F_1 = \frac{S_A/(s-1)}{S_c/st(r-1)} \sim F(s-1, st(r-1));$$

$$H_2 \text{ 成立时, } F_2 = \frac{S_B/(t-1)}{S_c/st(r-1)} \sim F(t-1, st(r-1));$$

$$H_3 \text{ 成立时, } F_3 = \frac{S_{A \times B}/(s-1)(t-1)}{S_c/st(r-1)} \sim F((s-1)(t-1), st(r-1)).$$

5.2 正交设计

- 有 $m \geq 2$ 个因素，i 个因素 F_i 有 s_i 个水平，m 较大时，难以安排全面试验。
- 正交设计研究可加模型，任意多个因素间不存在交互作用。
- 当因素 F_i 取 λ_i 时，可加模型为 $Y = \beta_0 + \sum_{i=1}^m \beta_i(\lambda_i) + e$ ，第一项称为一般平均，第

二项称为主效应，第三项随机误差服从正态分布。 $\sum_{\lambda_i=1}^{s_i} \beta_i(\lambda_i) = 0$ 。

• 定义 $A=(a_{ij})$ 是 $n \times m$ 矩阵，其第 j 列元素由 $1, 2, \dots, s_j$ 组成，如果对于任意 $j_1 < j_2$ ， $u \in (1, \dots, s_{j_1})$ ， $v \in (1, \dots, s_{j_2})$ ， $|\{i: (a_{ij_1}, a_{ij_2}) = (u, v)\}| = n/(s_{j_1} \times s_{j_2})$ ，则称 A 正交表。若 $s_1 = \dots = s_m = s$ ，记 A 为 $L_n(s^m)$ 。

6 序贯分析初步

6.1 序贯分析的重要性与两个要素

• 给定随机变量序列 $\{X_1, X_2, \dots\}$ ，称取值于 $\{1, 2, \dots\}$ 的随机变量 τ 为停时，有 $\{\tau = n\} = \{(X_1, \dots, X_n) \in B_n\}$ 。称 τ 是封闭的，如果 $P(\tau < \infty) = 1$ 。

- 考虑假设检验 $H_1: f=f_1 \leftrightarrow H_2: f=f_2$ ，似然比 $\lambda_n = \frac{\prod_{i=1}^n f_2(X_i)}{\prod_{i=1}^n f_1(X_i)}$ 。 λ_n 太小接受 f_1 ， λ_n

太大接受 f_2 , 否则再抽一个。取待定常数 $0 < A < 1 < B$ 。

- 序贯概率比检验(SPRT): $\tau = \inf\{n: \lambda_n \leq A \text{ 或 } \lambda_n \geq B\}$ 。 $\log(\lambda_n) = \sum_{i=1}^n \log\left(\frac{f_2(X_i)}{f_1(X_i)}\right)$ 是

随机游动。如果 $m(f_1 \neq f_2) > 0$, 则 τ 封闭。

- 设 a 是第一类错误概率, b 是第二类错误概率, 则 $a \leq \frac{1}{B}(1-b), b \leq A(1-a)$ 。实

际应用中, 给定 a, b , 通常取 $A = \frac{b}{1-a}, B = \frac{1-b}{a}$ 。

• 设 $\Delta' = (\tau', d')$ 为任意序贯检验法, 其两类错误概率 $a' \leq a, b' \leq b$, 其中 a, b 是 SPRT 两类错误的概率, 则 $E_i \tau \leq E_i \tau'$ 。【控制错误率, 最早停止】

7 统计决策与贝叶斯统计大意

7.1 统计决策问题概述

• 称 $R(\theta, \delta) = E_X L(\theta, \delta(X_1, \dots, X_n))$ 为决策 δ 的风险函数。最优决策: $R(\theta, \delta^*) \leq R(\theta, \delta)$, 对任意 $\theta \in \Theta$ 成立。

• 称决策 δ 是可容许的, 如果不存在 δ' , 使得 $R(\theta, \delta') \leq R(\theta, \delta)$, 且存在 θ_0 使得不等号严格成立。

- 决策 δ^* 是 minimax 决策, 若任意 δ , 有 $\sup\{R(\theta, \delta^*) | \theta \in \Theta\} \leq \sup\{R(\theta, \delta) | \theta \in \Theta\}$ 。

7.2 什么是贝叶斯统计

- 设 θ 的先验分布为 $\pi(\theta)$, 决策为 δ , 记 $\rho(\delta) = \int R(\theta, \delta) \pi(\theta) d\theta$ 为 δ 的平均风险。

• 若决策 δ^* 使得平均风险达到最小, 则称 δ^* 为贝叶斯决策。

- 得到数据 X 后, 只需在行动空间找 δ 使 $\int_0 L(\theta, \delta(x)) \pi(\theta | x) d\theta$ 达到最小。

7.3 关于先验分布

• 如果先验分布的类型使得后验分布仍为此类型, 则称先验分布是密度函数的共轭分布。

- 设 X_1, \dots, X_n 是来自两点分布 $B(1, p)$ 的简单随机样本, 若取参数 p 的先验分布为

$\text{Beta}(a, b)$, 则 p 的后验分布为 $\text{Beta}(a + \sum_{i=1}^n x_i, b + n - \sum_{i=1}^n x_i)$

- 设 X_1, \dots, X_n 是来自 Poisson 分布的简单随机样本, 取参数 λ 的先验分布为 $\Gamma(a, b)$,

则 λ 的后验分布为 $\Gamma(a + \sum_{i=1}^n x_i, b + n)$ 。

- 伽马分布是指数分布的共轭分布, 正态分布是正态分布的共轭分布。

• 称 Y 服从逆伽马分布, 如果 $1/Y$ 服从伽马分布, 记为 $\text{IG}(a, b)$ 。

• 正态分布均值已知, 取方差的先验分布为逆伽马分布, 则后验分布仍为逆伽马分布。

- 均匀分布的共轭分布是 Pareto 分布。

• 层次贝叶斯学派: 参数 θ 服从带有超参数 α 的分布, 而 α 又服从某个分布。

7.4 马氏链随机模拟

- 利用随机模拟方法，通过分步抽样，构造一个适当的马氏链，得到近似的、相依的后验分布模拟样本，并通过此样本计算后验的特征，如均值、分位数等。
- Gibbs sampler: $\mathbf{a}=(a_1, \cdots, a_n)$ ，固定其他分量 $\mathbf{a}_{-i}=(a_1, \cdots, a_{i-1}, a_{i+1}, \cdots, a_n)$ ， $p(a_i|\mathbf{a}_{-i})$ 往往很简单。
- MH 算法：根据 $p(\mathbf{a}^{(m)}, \mathbf{a}^{(m+1)})$ 选择 $\mathbf{a}^{(m+1)}$ ，再以 $\min\{1, r(\mathbf{a}^{(m)}, \mathbf{a}^{(m+1)})\}$ 概率接受。
- 估计方法：burn-in；采取若干次循环，每次循环只记录一个样本；独立产生若干个链，得到近似 i.i.d. 的样本。

8 抽样调查概述

8.1 简介

- 调查吸毒率 r ，可以问你是否出生在下半年且不吸毒？回答是的概率是 $(1-r)/2$ ，否的概率是 $(1+r)/2$ ，用 $n_1/n_2=(1-r)/(1+r)$ 计算出 r 。