

Lecture 9

It is better to have an approximate answer to the right question,

引言:

数据统计: $\begin{cases} \text{数据收集} \\ \text{数据分析} \\ \text{统计推断} \end{cases}$

e.g.

概念: 总体 (Population)

(回)

统计总体: 一个概率分布

Remark. ① 无限总体

② 有限总体 (离散)

个体数量很多时近似视为无限总体

③ 虚拟总体

统计模型: 一族概率分布

e.g. 正态分布族

指数分布族

参数模型: e.g. $\{N(\mu, \sigma^2)\}$

非参数模型: e.g. 连续型随机变量, $E(X)$ 存在, $Var(X)$ 有限
不能用有限个参数来刻画
直观上, 非参模型有很多参数

样本: $(X_1, \dots, X_n)$, 其中每个 X_i 来自总体
 n 称为样本容量

获取方式 $\begin{cases} \text{观测: 完全} \\ \text{不完全 (e.g. 电灯的使用寿命)} \\ \text{试验 (e.g. 抽样调查)} \end{cases}$

简单随机抽样: 总体个数 N 有限, 无放回

任意容量相等的样本都有相同的发生概率

随机样本: $X_1, \dots, X_n$ i.i.d., 即有放回或近似有放回

Remark. 不当抽样

e.g. 民意调查 { 有偏取样
无响应
问题设置: 如诱导性提问
语言
调查员的操作

统计量: $T(X_1, \dots, X_n)$ 样本的函数, 完全由样本决定, 是数据的简化方式

e.g. $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$,

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

$\bar{X} - \mu$ { μ 未知: 由于 μ 取决于总体, $\bar{X}$ 不是统计量
 μ 已知: 是统计量, 样本的函数

统计推断: 总体 $\xrightleftharpoons[\text{推断}]{\text{决定}}$ 样本(分布)

{ 经典方法: 频率解释
Bayes 方法

e.g. 元件寿命

• 假设 $X \sim \text{Exp}(\lambda)$, 问 $\lambda = ?$ (参数估计)

• 假设合格标准是 $E(X) \geq L \rightarrow$ 检验标准: $\bar{X} \geq l$ (假设检验)

e.g. 设模型 $Y_i = aX_i + \varepsilon_i$

• $a = ?$ ($X_i, Y_i, i \in \{1, \dots, n\} \Rightarrow a$)

• $X_i = ?$ (a, Y_i 已知)

(模型推断, 参数估计)
(变量推断)

Chapter 6 参数估计

1° 矩估计

设 $X_1, \dots, X_n$ iid

样本矩: k 阶原点矩 $a_k \triangleq \frac{1}{n} \sum_{i=1}^n X_i^k \xrightarrow{\text{LLN}} E(X^k)$

k 阶中心矩 $m_k \triangleq \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^k \xrightarrow{\text{LLN}} E((X - \mu)^k)$

e.g. $X_i \text{ iid } N(\mu, \sigma^2)$

$$\mu = E(X) \approx a_1 = \bar{X}$$

$$\sigma^2 = \text{Var}(X) \approx m_2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

e.g. $X_i \text{ iid } \text{Exp}(\lambda)$

$$\frac{1}{\lambda} = E(X) \approx a_1 \Rightarrow \lambda = \frac{1}{a_1}$$

$$\frac{1}{\lambda^2} = \text{Var}(X) \approx m_2 \Rightarrow \lambda = \left(\frac{1}{m_2}\right)^{\frac{1}{2}}$$

基本原则: 尽量用低阶矩 (计算稳定性好)

e.g. 经验分布 $F_n(x) = \frac{\{x_1, \dots, x_n \text{ 中不超过 } x \text{ 的个数}\}}{n}$

$$= \frac{1}{n} \sum_{i=1}^n I_i(x), \quad I_i(x) = \begin{cases} 1, & x_i \leq x \\ 0, & x_i > x \end{cases}$$

$$I_i(x) \stackrel{iid}{\sim} B(F(x))$$

$$\Rightarrow F_n(x) \xrightarrow{a.s.} F(x)$$

(?) $F_n(x)$ — 在每个数据点 x_i 上的概率为 $\frac{1}{n}$ 的 cdf.

(?) Remark. $F_n(x)$ 的矩即 $(x_1, \dots, x_n)$ 的样本矩.

2° 极大似然估计 (MLE)

假设 $x_1, \dots, x_n$ 的联合分布为 $f(x_1, \dots, x_n; \theta)$ (pdf/pmf)

θ 为参数, 可为向量或标量

对于观测 $(x_1, \dots, x_n)$ 的似然 (函数) 定义为:

$$L(\theta) = f(x_1, \dots, x_n; \theta)$$

Remark. ① 观测数据常记为 $x_1, \dots, x_n$, 视为 $x_1, \dots, x_n$ 的一个实现

② 在离散情形 $L(\theta)$ 是出现 $x_1, \dots, x_n$ 的概率

③ 若 $x_1, \dots, x_n$ iid; 总体分布记为 $f_1(x; \theta)$

$$L(\theta) = f(x_1, \dots, x_n; \theta) = f_1(x_1; \theta) \cdots f_1(x_n; \theta)$$

e.g. $x_i \stackrel{iid}{\sim} N(\mu, \sigma^2)$, μ, σ^2 未知

$$L(\mu, \sigma^2) = \prod_{i=1}^n \left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}} \right)$$

定义: $\theta^* \triangleq \arg \max_{\theta} L(\theta)$, 称为 θ 的极大似然估计

Remark. θ^* 随 $x_1, \dots, x_n$

e.g. 接上例

似然方程组.
$$\begin{cases} \frac{\partial \ln L}{\partial \mu} = 0, \\ \frac{\partial \ln L}{\partial \sigma^2} = 0. \end{cases} \Rightarrow \begin{cases} \mu^* = \bar{x} \\ (\sigma^2)^* = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \end{cases}$$

经验验证: $(\mu^*, (\sigma^2)^*)$ 为 $\ln L$, 也就是 L 的最大值点

Remark ① 上述结果与矩估计结果一致, 这是个案

② MLE 的不变性: $g(\theta^*) = g(\theta^*)$

e.g. 设 $X_i \text{ iid } U(0, \theta)$. θ 未知

$$L(\theta) = \begin{cases} \frac{1}{\theta^n}, & (\forall i), X_i \in (0, \theta) \\ 0, & (\exists i), X_i \notin (0, \theta) \end{cases}$$

$$\Rightarrow \theta^* = \max \{X_1, \dots, X_n\}$$

由 $L(\theta^*) \geq L(\theta)$,

有 θ^* 为 θ 的 MLE.

e.g. 设总体 pdf 是 $f(x; \theta) = \frac{1}{\pi(1+(x-\theta)^2)}$, $x \in \mathbb{R}$ (Cauchy 分布)

① 没有矩, 不能进行矩估计

② MLE: 似然方程为 $\sum_{i=1}^n \frac{X_i - \theta}{1 + (X_i - \theta)^2} = 0$

不易求解

③ θ 为中位数

$\theta \approx$ 样本中位数

Remark. ① 估计方法不唯一, 没有绝对优劣

② MLE 方程的根不都是解

③ MLE 必须先有 f 的表达式

矩估计只需参数与矩的关系

④ 似然在极大值点处平滑不易估计

Lecture 10

Remark $\left. \begin{array}{l} \text{参数估计} \\ \text{MLE} \end{array} \right\} \text{点估计}$

经典方法: 参数 θ 被视为数或数组.

3° Bayes 估计

在搜集数据前, 对参数 θ 有先验知识, 可用一个概率分布刻画, 称为先验分布

用 Θ 来表示 θ 分布的随机变量.

$f_{\Theta}(\theta)$ —— 先验分布 (主观)

X —— 试验

$f_X(x|\theta)$ —— 样本分布

$$f_{\Theta}(\theta|x) = \frac{f(x, \theta)}{f_X(x)} = \frac{f_X(x|\theta)f_{\Theta}(\theta)}{f_X(x)} \text{ —— 后验分布}$$

e.g. 正面向上的概率为 θ , 抛 n 次, 出现了 x 次正面向上

先对 θ 的初始认知状态进行刻画:

$$\Theta \sim U(0,1) \text{ i.e. } f_{\Theta}(\theta) = 1, \theta \in (0,1) \text{ (同等无知原则, Bayes 原则)}$$

记 x : 抛 n 次, 正面向上的次数, 对给定 θ :

$$x \sim B(n, \theta) \text{ i.e. } f_x(x|\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}, x \in [0, n] \cap \mathbb{N}$$

更新 θ 的分布, 考虑 (x, Θ) 的联合分布

$$f(\theta, x) = f_{\Theta}(\theta) \cdot f_x(x|\theta) = 1 \cdot \binom{n}{x} \theta^x (1-\theta)^{n-x}, x \in [0, n] \cap \mathbb{N}, \theta \in (0,1)$$

连续 离散

故

$$f_x(x) = \int_0^1 f(\theta, x) d\theta$$

$$\text{由 } \int_0^1 u^{a-1} (1-u)^{b-1} du = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}, \text{ 其中 } \Gamma(x) = \int_0^{+\infty} t^{x-1} e^{-t} dt$$

$$\text{有 } f_x(x) = \frac{1}{n+1}$$

$$\Rightarrow f_{\Theta}(\theta|x) = \frac{\Gamma(n+2)}{\Gamma(x+1)\Gamma(n-x+1)} \theta^x (1-\theta)^{n-x}$$

$$\sim B(x+1, n-x+1)$$

Remark. 先验分布是 $B(1,1)$

可用后验分布期望 $E(\Theta|x)$ 或后验分布众数 $f(\theta|x)$ 的最大值来估计 θ .

Remark. Bayes. $\hat{\theta} = E(\Theta|x) = \frac{x+1}{n+2}$ (在 n 很小时, 估计不极端)

$$\text{MLE. } \theta^* = \frac{x}{n}$$

Remark. 若先验分布是 $B(a,b)$, 则后验分布是 $B(a+x, b+n-x)$

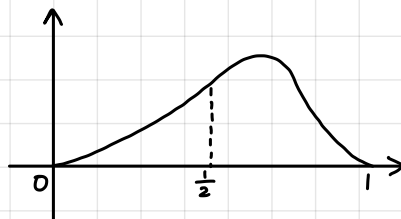
$$\hat{\theta} = E(\Theta|x) = \frac{a+x}{a+b+n} = \frac{a+b}{a+b+n} \cdot \frac{a}{a+b} + \frac{n}{a+b+n} \cdot \frac{x}{n}$$

$B(a,b)$ 的均值 $B(x, n-x)$ 的均值

e.g. $n=20, x=13$, 则后验分布 $B(14,8)$, 如图

$\theta < \frac{1}{4}$: 不太可能

而 $P(\theta > \frac{1}{2}) \approx 0.91$



Remark. 先验分布的选取

4° 无偏性.

偏差 (Bias): $E_0(\hat{\theta} - \theta) = E_0(\hat{\theta}) - \theta$, θ 未知, 但在这里当做一个确定数

定义: 若 $E_0(\hat{\theta}) = \theta$ 对 θ 的任意可能值都成立, 则称 $\hat{\theta}$ 为 θ 的一个无偏估计

Remark. 一般地, 称 $\hat{g}(x_1, \dots, x_n)$ 是 $g(\theta)$ 的一个无偏估计.

$$\Leftrightarrow E(\hat{g}(x_1, \dots, x_n)) = g(\theta).$$

Remark. ① 无偏即无系统误差.

② 若无偏, 由 LLN: $\frac{1}{N} \sum_{m=1}^N \hat{g}(x_1^{(m)}, x_2^{(m)}, \dots, x_n^{(m)}) \xrightarrow{a.s.} E(\hat{g}) = E(g(\theta))$

③ 无偏性的实际价值视情况而定

e.g. $E(X) = \mu, \text{Var}(X) = \sigma^2$

$$\Rightarrow E(\bar{X}) = \mu, E(m_2) = E\left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{X})^2\right) = \frac{n-1}{n} \sigma^2 \text{ (系统性偏小)}$$

而 $S^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{X})^2$ 是无偏的.

Remark. 注意区别总体方差和样本方差.

e.g. 设 $x_i \sim U(0, \theta)$, $\text{MLE: } \theta^* = \max\{x_1, \dots, x_n\}$, $E(\theta^*) = \frac{n}{n+1} \theta$, 偏小.

Remark. ① MLE 不保证无偏

② $\frac{n+1}{n} \theta^*, 2\bar{X}$ 无偏

5° 均方误差准则 (有效性准则)

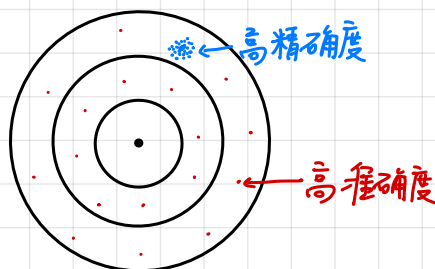
均方误差: $E_0((\hat{\theta} - \theta)^2)$

$$\Leftrightarrow \text{Var}(\hat{\theta} - \theta) + E_0^2(\hat{\theta} - \theta)$$

利用方差
的计算式:
 $\text{Var}(X) = E(X^2) - E^2(X)$

$$= \text{Var}(\hat{\theta}) + E_0^2(\hat{\theta} - \theta)$$

精确度
precision
数据扰动
造成的影
准确度
accuracy
算法能力



定义 (均方误差准则) 设 $\hat{\theta}_1, \hat{\theta}_2$ 都是 θ 的无偏估计

若 $\text{Var}(\hat{\theta}_1) \leq \text{Var}(\hat{\theta}_2)$ 对于 $\forall \theta$ 成立且 $\exists \theta_0$ s.t. $\text{Var}_{\theta_0}(\hat{\theta}_1) < \text{Var}_{\theta_0}(\hat{\theta}_2)$

则称估计 $\hat{\theta}_1$ 在均方误差意义下优于 $\hat{\theta}_2$.

e.g. 估计 μ .

$$\mu = E(\bar{X}) = E(X_1) = E\left(\frac{1}{2}(X_1 + X_2)\right)$$

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

$$\text{Var}(X_1) = \sigma^2$$

$$\text{Var}\left(\frac{1}{2}(X_1 + X_2)\right) = \frac{\sigma^2}{2}.$$

故 $\bar{X}$ 是最优的.

定义 设 $\hat{\theta}_0$ 是 θ 的无偏估计, 且 $\text{Var}(\hat{\theta}_0) \leq \text{Var}(\hat{\theta})$ 对 $\forall$ 无偏估计 $\hat{\theta}$ 成立, 则称 $\hat{\theta}_0$ 是 θ 的最小方差无偏估计 (MVU估计)

e.g. $X \sim N(\mu, \sigma^2)$

$$E(m_2) = \frac{n-1}{n} \sigma^2, \quad E(s^2) = \sigma^2$$

注意到 $E((m_2 - \sigma^2)^2) < E((s^2 - \sigma^2)^2)$

可以用低偏换方差更小、

6° 大样本性质

$\hat{\theta} = \hat{\theta}(x_1, \dots, x_n)$, 当 $n \rightarrow \infty$ 时的性质

· 渐近无偏: $\lim_{n \rightarrow \infty} E(\hat{\theta} - \theta) = 0$

· 相合性: 若 $\forall \varepsilon > 0$, 有 $\lim_{n \rightarrow \infty} P(|\hat{\theta} - \theta| \geq \varepsilon) = 0$, 则称 $\hat{\theta}$ 是 θ 的一个相合估计 (i.e. $\hat{\theta} \xrightarrow{P} \theta$)

Remark. 相合性是良好估计的自然要求

e.g. $m_2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$ (中插一个 μ)

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

$$\begin{array}{c} \downarrow P \\ \sigma^2 \\ \Rightarrow m_2 \xrightarrow{P} \sigma^2 \end{array}$$

$$\downarrow P \\ 0$$

收敛吗?

将 LLN 和 CLT 指出的 $\bar{x}$ 的优良性质推广到一般的 $\hat{\theta}$

· 渐近正态性: 若存在 σ_n s.t. ① $\lim_{n \rightarrow \infty} \sigma_n = 0$

$$\textcircled{2} \lim_{n \rightarrow \infty} P\left(\frac{\hat{\theta} - \theta}{\sigma_n} \leq x\right) = \Phi(x)$$

则称 $\hat{\theta}$ 为 θ 的一个相合渐近正态估计

Remark. ① 由 CLT: $\bar{x}$ 是 μ 的一个相合渐近正态估计

② n 很大时, 可用 $N(\theta, \sigma_n^2)$ 来近似 $\hat{\theta}$ 的分布

③ 有时, 取 $\sigma_n^2 = \text{Var}(\sigma)$

7° MLE 的优良性.

$L(\theta) = f(x_1, \dots, x_n; \theta)$ 提取了样本中所有关于 θ 的信息

$$\theta^* = \arg \max_{\theta} L(\theta)$$

① 不变性

② 相合性

③ 渐近正态

④ 渐近有效性

⑤ 接近 Bayes 估计

看第③条: $\frac{\theta^* - \theta}{\sigma_n} \rightarrow N(0, 1)$

Fisher 信息量:

$$I(\theta) \triangleq E\left(\left(\frac{\partial \ln f(x; \theta)}{\partial \theta}\right)^2\right)$$

$$= \text{Var}\left(\frac{\partial \ln f(x; \theta)}{\partial \theta}\right)$$

$$= -E\left(\frac{\partial^2 \ln f(x; \theta)}{\partial \theta^2}\right)$$

$$I_n(\theta) \triangleq E\left(\left(\frac{\partial \ln f(x_1, \dots, x_n; \theta)}{\partial \theta}\right)^2\right)$$

$$\stackrel{\text{iid}}{=} n I(\theta)$$

对 σ_n^2 取 $\frac{1}{I_n(\theta)}$ 或 $\frac{1}{I_n(\theta^*)}$

$$\frac{\partial \ln f(x; \theta)}{\partial \theta} = \frac{f'_\theta}{f}$$

$$E\left(\frac{\partial \ln f(x; \theta)}{\partial \theta}\right) = \int_{\mathbb{R}} \frac{f'_\theta}{f} \cdot f dx$$

$$= \int_{\mathbb{R}} f'_\theta dx \equiv 0$$

积分求导交换

$$\left(\frac{f'_\theta}{f}\right)'_\theta = \frac{f''_{\theta\theta}}{f} - \frac{(f'_\theta)^2}{f^2}$$

两边期望

Lecture 11

Review & Remarks.

1. $x_1, \dots, x_n$: 样本 ("随机样本" 通常指 iid)

样本分布: $f(x_1, \dots, x_n; \theta)$

$$\hat{\theta} = \hat{\theta}(x_1, \dots, x_n)$$

$$\text{标准误差: } se(\hat{\theta}) = \sqrt{\text{Var}(\hat{\theta})}$$

指统计量的标准差.

$$\text{标准误差的估计: } \hat{se} = \hat{se}(\hat{\theta})$$

2. 经典点估计的优良性.

$$n \text{ 固定: } \begin{cases} \text{无偏性: } E(\hat{\theta} - \theta) = 0 \\ \text{有效性: } \text{MSE: } E[(\hat{\theta} - \theta)^2] = \text{Var}(\hat{\theta}) + E^2(\hat{\theta} - \theta) \end{cases}$$

$$\text{Var}(\theta) = \text{Var}(\hat{\theta} - \theta)$$

↑
常数

3. Bayes 估计:

· 后验众数与 MLE 的关系.

设 $f_{\theta}(x) \propto 1$ (同等无知原则)

$$f_{\theta}(x) = \frac{f_{\theta}(x) \cdot f(x|\theta)}{f_{\theta}(x)} \propto f(x|\theta) = f(x; \theta) = L(\theta)$$

故后验众数 = MLE 估计值.

4. 统计决策与损失函数

损失函数 $L(\theta, \hat{\theta})$

$\hat{\theta}$ 的风险: $R(\theta, \hat{\theta}) \triangleq E(L(\theta, \hat{\theta}))$

$$\stackrel{\text{连续时}}{=} \int L(\theta, \hat{\theta}) f(x_1, \dots, x_n; \theta) dx_1 \dots dx_n.$$

Bayers 风险: $r(\theta, \hat{\theta}) \triangleq \int R(\theta, \hat{\theta}) f_{(H)}(\theta) d\theta.$

将 θ 视为分布 (H) 的一个实现.

$$= \int L(\theta, \hat{\theta}) f_{x_1, \dots, x_n}(\theta) f_{(H)}(\theta) dx_1 \dots dx_n$$

eg. $L(\theta, \hat{\theta}) = (\hat{\theta} - \theta)^2.$

$$R(\theta, \hat{\theta}) = E[(\hat{\theta} - \theta)^2] \quad \text{i.e. MSE}$$

$$r(\theta, \hat{\theta}) = E[(\hat{\theta} - \theta)^2]$$

End Review & Remarks.

8. 区间估计

定义: 给定 $\alpha \in (0, 1)$, 对 θ 的任意可能取值, 有

$$P(\hat{\theta}_1 < \theta < \hat{\theta}_2) \geq 1 - \alpha.$$

则称 $(\hat{\theta}_1, \hat{\theta}_2)$ 为 θ 的 $(1 - \alpha)$ 置信的区间估计

Remark.

① α 常取 0.05, 0.01, 0.1

② 可靠度优先 i.e. 优先保证 $P(\hat{\theta}_1 < \theta < \hat{\theta}_2) \geq 1 - \alpha.$
再尽量提高精度 $(\hat{\theta}_2 - \hat{\theta}_1).$

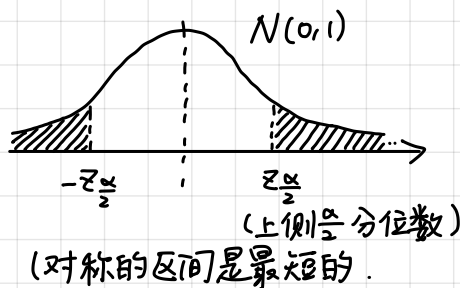
eg. $\bar{X} \sim N(\mu, \sigma^2), \sigma^2$ 已知, μ 未知.

欲使 $P(\bar{X} - c_1 < \mu < \bar{X} + c_2) \geq 1 - \alpha$

$$\text{即 } P(-c_2 \leq \bar{X} - \mu < c_1) \geq 1 - \alpha$$

而 $\bar{X} - \mu \sim N(0, \frac{\sigma^2}{n})$

$$\Rightarrow \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$



$\Rightarrow$ 所求的区间估计为

$$(\bar{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}})$$

Remark. ① 有时也记为 $\bar{X} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$

② 若 $\alpha = 0.05$, 有 $z_{\frac{\alpha}{2}} \approx 1.96$.
有时也取 $z_{\frac{\alpha}{2}} = 2$.

Remark. ① 用 $\bar{x}$ 估计 μ , 则有 $(1-\alpha)$ 的置信使得所得估差 $|\bar{x}-\mu| \leq z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$.

② $\frac{\sigma}{\sqrt{n}}$ —— ($\bar{x}$ 的标准差) 标准误差.

③ 给定 $\varepsilon > 0$, $n \sim (\frac{z_{\frac{\alpha}{2}} \sigma}{\varepsilon})^2$ 时, 有 $(1-\alpha)$ 置信. 误差 $\leq \varepsilon$.

e.g. 设 $\bar{x} \sim N(\mu, \sigma^2)$, μ, σ^2 未知. 估计 μ

$$\text{则 } \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}} \sim t(n-1)$$

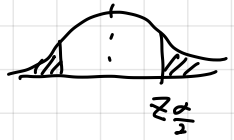
$$\text{故所求区间 } (\bar{x} - t_{\frac{\alpha}{2}}(n-1) \frac{s}{\sqrt{n}}, \bar{x} + t_{\frac{\alpha}{2}}(n-1) \frac{s}{\sqrt{n}})$$

$$\frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}} \sim N(0,1)$$

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$$

$$|\bar{x}-\mu| \leq \varepsilon, 1-\alpha$$

$$\left| \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}} \right| \leq \frac{\varepsilon \sqrt{n}}{s} = z_{\frac{\alpha}{2}}$$



$$X \sim N(0,1)$$

$$Y \sim \chi^2(n)$$

$$Z = \frac{X}{\sqrt{\frac{Y}{n}}}$$

$$Z \sim t(n)$$

Remark. ① 正态假设下, s^2 与 $\bar{x}$ 独立.

$$\textcircled{2} \frac{(n-1)s^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \bar{x})^2 \xrightarrow{\text{插}\mu} \sum_{i=1}^n \left(\frac{x_i - \mu}{\frac{s}{\sqrt{n}}} \right)^2 - \left(\frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \right)^2$$

$$\sim \chi^2(n-1)$$

枢轴变量法:

(1) 找 θ 的相关统计量 $\hat{\theta}(x_1, \dots, x_n)$ (通常是 θ 的良好点估计)

(2) 找出函数 $H(\hat{\theta}, \theta)$ 的分布 (与未知变量无关)

枢轴变量.

(3) 求 $(\hat{\theta}_1, \hat{\theta}_2)$

用样本统计量估计总体参数.
估计值是数轴上一点.

求枢轴变量的分位数

求出相应统计量的置信区间

因此, 本课程只求理解出现过的几个.

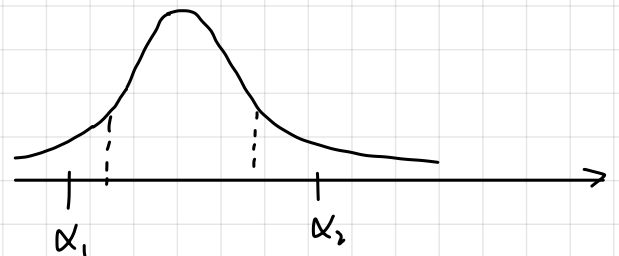
e.g. $\bar{x} \sim (\mu, \sigma^2)$, μ, σ^2 未知, 估计 σ^2 .

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$$

$\Rightarrow$ 所求区间

$$\left(\frac{(n-1)s^2}{\chi_{\frac{\alpha}{2}}^2(n-1)}, \frac{(n-1)s^2}{\chi_{1-\frac{\alpha}{2}}^2(n-1)} \right)$$

直接用 χ^2 的分位数



$$\text{取 } \alpha_1 = \alpha_2 = \frac{\alpha}{2}$$

而不考虑最小化区间长度.

e.g. $\bar{x} \sim N(\mu_1, \sigma^2)$, $\bar{y} \sim N(\mu_2, \sigma^2)$ 独立, μ_1, μ_2, σ^2 未知, 估计 $(\mu_1 - \mu_2)$.

$x_1, \dots, x_n; y_1, \dots, y_m$ 为随机样本.

$$(\bar{x} - \bar{y}) - (\mu_1 - \mu_2) \sim N(0, \frac{\sigma^2}{n} + \frac{\sigma^2}{m}) \quad // \quad \hat{\sigma}^2 = \frac{\sigma^2}{n} + \frac{\sigma^2}{m}$$

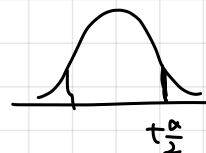
$$S^2 \triangleq \frac{n-1}{n+m-2} S_1^2 + \frac{m-1}{n+m-2} S_2^2$$

$$= \frac{1}{n+m-2} \left(\sum_{i=1}^n (x_i - \bar{x})^2 + \sum_{j=1}^m (y_j - \bar{y})^2 \right)$$

$$\left| \frac{(n-1)S_1^2}{\sigma^2} + \frac{(m-1)S_2^2}{\sigma^2} \right| \sim \chi^2(n+m-2).$$

$$E_p \left(\frac{n+m-2}{\sigma^2} S^2 \right) \sim \chi^2(n+m-2)$$

$$\frac{\frac{(\bar{x} - \bar{y}) - (\mu_1 - \mu_2)}{\sigma'}}{\sqrt{\frac{(n+m-2)S^2}{\sigma^2(n+m-2)}}} \sim t(n+m-2) \quad \text{即} \quad \frac{(\bar{x} - \bar{y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{1}{n} + \frac{1}{m}} \cdot S} \sim t(n+m-2)$$



渐近置信区间 (大样本方法)

e.g. (选举问题), 记事实支持率是 p , $n=1200$, 观测值 $\frac{684}{1200} = 0.57$
给出一个 95% 置信的区间估计

考虑 $x_i \sim B(p)$ (近似 iid). $P_n \triangleq \bar{x}$.

$$\mu = p, \sigma^2 = p(1-p)$$

$$E(P_n) = p, \text{Var}(P_n) = \frac{p(1-p)}{n} \Rightarrow \text{se}(P_n) = \sqrt{\frac{p(1-p)}{n}}$$

在这里
 P_n 是随机变量,
不是观测值

$$\frac{P_n - p}{\text{se}(P_n)} \xrightarrow{\text{CLT}} N(0,1)$$

标准差与 p 有关,
这造成估计上的
困难.

$$\text{要使 } P\left(\frac{|P_n - p|}{\sqrt{\frac{p(1-p)}{n}}} < z_{\frac{\alpha}{2}}\right) \geq 1 - \alpha.$$

$$\textcircled{1} (P_n - p)^2 < z_{\frac{\alpha}{2}}^2 \frac{p(1-p)}{n} \Rightarrow P_{\pm} = \frac{1}{1 \pm \frac{z_{\frac{\alpha}{2}}^2}{n}} \left(P_n + \frac{z_{\frac{\alpha}{2}}^2}{2n} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{P_n(1-P_n)}{n} + \frac{z_{\frac{\alpha}{2}}^2}{4n}} \right)$$

取 $\alpha = 0.05$
得区间 $(0.542, 0.598)$

$$\textcircled{2} \text{ 用 } S^2 \text{ 来近似 } \sigma^2 = p(1-p), \text{ 近似认为 } \frac{P_n - p}{\frac{S}{\sqrt{n}}} \sim N(0,1)$$

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = 0.2475$$

$$\text{取 } \alpha = 0.05$$

$$\text{得区间 } (0.542, 0.598)$$

样本给完后,
 P_n 即已知,
而 p 才是不知道
的量.

$$\textcircled{3} \text{ 用 } m_2 \text{ 来近似 } \sigma^2.$$

这等价于用 $P_n(1-P_n)$ 近似 σ^2

$$\text{取 } \alpha = 0.05$$

$$\text{得区间 } (0.542, 0.598)$$

$$\textcircled{4} \text{ 用 } \frac{1}{4} \text{ 来近似 } \sigma^2 = p(1-p). \text{ 取 } \sigma^2 \text{ 最大值.}$$

$$\Rightarrow (0.542, 0.599)$$

① 置信 $\sim 1-\alpha$ 时, 近似程度取决于 n 的大小, 及总体分布

② $X_i \stackrel{iid}{\sim} B(p)$. $p^* = P_n$. $f(x_i; p) = p^{x_i} (1-p)^{1-x_i}$ 第 i 个变量的 pdf., $I(p) = \frac{1}{p(1-p)}$ $B(p)$ 的 pdf.

$$I_n(p) = n I(p) = \frac{n}{p(1-p)}$$

$$\text{取 } \sigma_n = \sqrt{\frac{1}{I_n(p)}} = \sqrt{\frac{p(1-p)}{n}} \Rightarrow \frac{P_n - p}{\sqrt{\frac{p(1-p)}{n}}} \xrightarrow{\text{近似}} N(0,1).$$

$$\text{取 } \sigma_n = \sqrt{\frac{1}{I_n(p^*)}} = \sqrt{\frac{P_n(1-P_n)}{n}} \Rightarrow \frac{P_n - p}{\sqrt{\frac{P_n(1-P_n)}{n}}} \xrightarrow{\text{近似}} N(0,1)$$

(这里 MLE 与 P_n 结果恰一致)

MLE 渐进正态性

e.g. $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$, 独立, $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$ 未知, 估计 $\mu_1 - \mu_2$
随机样本 $X_1, \dots, X_n, Y_1, \dots, Y_m$.

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \sim N(0,1)$$

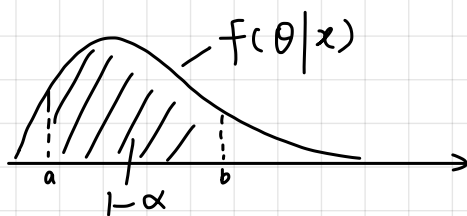
用样本方差来近似.

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n} + \frac{S_2^2}{m}}} \xrightarrow{\text{近似}} N(0,1)$$

Bayes 区间估计

有 θ 后验分布 $f_\theta(\theta|x)$

$$P(a < \theta < b | x) \geq 1 - \alpha$$



最大后验区间估计 $f_{\theta^+}(\theta|x) \geq f_{\theta^-}(\theta|x)$

$$\forall \theta_1 \in (a,b), \forall \theta_2 \notin (a,b)$$

e.g. $X \sim N(\mu, \sigma^2)$, σ^2 已知.

取 $f(\mu) \propto 1$, 后验分布 $N(\bar{X}, \frac{\sigma^2}{n})$

$$\Rightarrow \frac{\mu - \bar{X}}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1) \Rightarrow (\bar{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}})$$

Lecture 12

Review: 1° 置信区间 v.s. Bayes 区间

2° 小样本方法 v.s. 大样本方法

{ 小样本 — 精确分布
大样本 — 近似分布 (多依赖于 CLT, 是渐近置信区间.)

枢轴变量 { 良好点估计
概率分布可用

Chapter 7 假设检验

1° 基本概念

e.g. 一大批电子元件寿命 X , 样本 $X_1, \dots, X_n$ iid.

Q_1 : 若假设 $X \sim \text{Exp}(\lambda)$, 问 $\lambda = ?$

这是参数估计 (已知分布族)

Q_2 : 设合格标准: $\mu = E(X) \geq 5000$, 如何判定是否合格?

尝试建立执行标准 $\bar{X} \geq L$. 那么 $L = ?$

→ 假设检验: 样本在多大程度上支持假设

定义: 统计假设: 对一个或多个总体的某种推断或猜测

原假设 (null hypothesis) H_0 : 被检验假设

备择假设 (alternative hypothesis) H_1 : 拒绝 H_0 后可供选择的假设
也叫研究假设

Remark. 若假设可以表示成参数形式, 那么 H_0 可表示为 $H_0: \theta \in \Theta_0$; $H_1: \theta \in \Theta_1$
且 $\Theta_0 \cap \Theta_1 = \emptyset$, $\Theta_0 \cup \Theta_1 = \Theta$ 所有可能取值.

e.g. (1) $H_0: \lambda \leq \frac{1}{5000}$: $H_1: \lambda > \frac{1}{5000}$ → 对应于 λ 不同取值的无数个总体

双边检验

简单假设 — (2) $H_1: \mu = \mu_0$
对应单个总体

: $H_1: \mu \neq \mu_0$
(已知 $N(\mu, \sigma^2)$, σ^2 已知)

单边检验

(3) $H_1: \mu_1 = \mu_2$

: $H_2: \mu_1 > \mu_2$

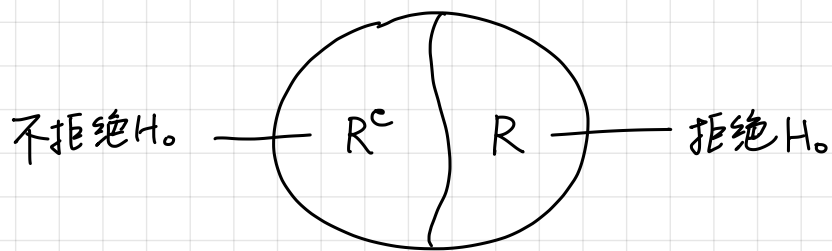
复合假设: 对应多个总体

拒绝 H_0 ; 不拒绝 H_0 .

假设检验: 依据样本的决策过程.

检验: 作出决策的一个具体法则

所有可能观测的集合 $\{(x_1, \dots, x_n)\}$



R: 临界域 (critical region)

形式上抽象成:

$$R = \{(x_1, \dots, x_n) \mid T(x_1, \dots, x_n) \geq c\}$$

----- 临界值.

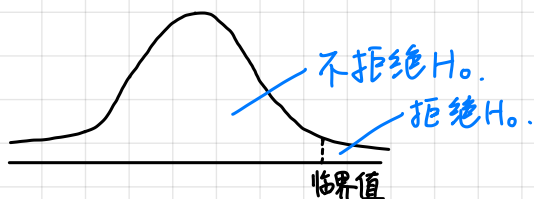
e.g. $\{(x_1, \dots, x_n) \mid \bar{x} < 1\}$

拒绝 H_0 : 在假设 H_0 为真的情况下, 所观测样本出现的概率

如果该概率很小, 意味着样本是拒绝 H_0 .

(Faith: 小概率事件不易发生)

e.g.



两种可能的错误

I类 (弃真错误)

II类 (纳伪错误)

	拒绝 H_0	不拒绝 H_0
H_0 真	X(I)	✓
H_0 假	✓	X(II)

对应的概率: $P_0(I), \theta \in \Theta_0$.

$P_0(II), \theta \in \Theta_1$, 注意, 这里是在 H_0 为假的前提下做条件概率.

Remark. ① 依样本决策, 错误不可避免

② 同一次检验不能同时犯两种错误

③ n 固定时, 两种错误发生的概率此消彼长

e.g. (直觉) (a) 都不拒绝: $P_0(I) = 0, P_0(II) = 1$ 反之类似
(b) ...

$$P(I) = P_0((x_1, \dots, x_n) \in R) \text{ when } \theta_0 \in \Theta_0, \text{ 记为 } \alpha(R)$$

$$P(II) = P_0((x_1, \dots, x_n) \in R^c) \text{ when } \theta_1 \in \Theta_1, \text{ 记为 } \beta(R)$$

对划分好的 R, α, β
是定义域不同的 θ 的函数.

// θ 不同 $\rightarrow$ 分布不同 $\rightarrow \alpha, \beta$ 变化

功效函数 (power function) 对不同给定的 θ , 给定划分 R , 拒绝 H 的概率.

$$P((X_1, \dots, X_n) \in R) = \begin{cases} \alpha(R) & , \theta \in \Theta_1 \\ 1 - \beta(R) & , \theta \in \Theta_2 \end{cases} \rightarrow \text{功效: } H_0 \text{ 为假时能将其拒绝.}$$

Neyman-Pearson 范式 (n 固定). 控制 $P_0(I) \leq \alpha$. $\forall \theta \in \Theta$. ($\alpha > 0$, 通常很小, 预先给定, 称为检验水平或显著水平)

再在这一限制下, 使 $P_0(I)$ 尽可能小.

Remark. α 固定, 使 $P_0(I)$ 最小的检验称为水平 α 下的一致最优检验 (不一定存在, 一般也不易求解)

Remark. H_0 与 H_1 , 一般地位不对等

① 原假设 H_0 通常是受保护的 // 而检验过程是在试图从样本中找证据拒绝 H_0 .

② 备则假设通常是在研究中感兴趣的.

2° 临界值检验法

e.g. $Z \sim (\mu, \sigma^2)$

(1) $H_0: \mu = \mu_0$, $H_1: \mu \neq \mu_0$.

(2) $H_0: \mu \geq \mu_0$, $H_1: \mu < \mu_0$.

(3) $H_0: \mu \leq \mu_0$, $H_1: \mu > \mu_0$.

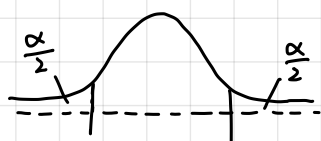
下面讨论 (1), (2).

Case I: 若 $\sigma^2 > 0$ 已知.

给定 α , 试图求解出划分 R (事实上就是参数 c)
s.t. $P(I \geq c) \leq \alpha$

(1) 给定 $\alpha > 0$, 用 Neyman-Pearson 范式, 当 H_0 为真时

控制 $P_0(I): P(|\bar{X} - \mu_0| \geq c) \leq \alpha$ (when H_0 为真).



此时: $\bar{X} - \mu_0 \sim N(0, \frac{\sigma^2}{n})$

$$\text{即 } P\left(\left|\frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}\right| \geq \frac{c}{\frac{\sigma}{\sqrt{n}}}\right) \leq \alpha$$

$\Rightarrow$ 检验(准则): $|\bar{X} - \mu_0| \geq z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$

即这样的划分保证了
在 RUR^c 中 $\forall$ 取样

基本思路:

$|\bar{X} - \mu_0| \geq c$ 时拒绝
用此事件给出概率.

$\longleftrightarrow$ 事件发生的概率落在
我所指定的
概率为 α 的域 R 中

(2) 给定 $\alpha > 0$.

当 H_0 为真时.

$$P(\bar{X} \leq c) \leq \alpha$$



给出模型

$$P\left(\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq \frac{c - \mu}{\frac{\sigma}{\sqrt{n}}}\right) \leq \alpha$$

↓
标准化

设 $Z \sim N(0, 1)$

$$P\left(Z \leq \frac{c - \mu}{\frac{\sigma}{\sqrt{n}}}\right) \leq \alpha$$

↓
用分位数表示

$$\Rightarrow \Phi\left(\frac{c - \mu}{\frac{\sigma}{\sqrt{n}}}\right) \leq \alpha \quad \text{对一切 } \mu \geq \mu_0 \text{ 都成立, 有}$$

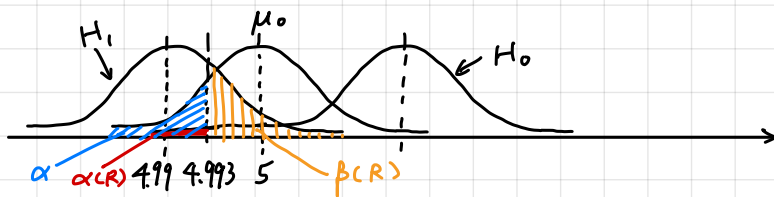
$$\max \frac{c - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{c - \mu_0}{\frac{\sigma}{\sqrt{n}}} \leq -z_\alpha \quad (\text{注意这里是下侧分位数})$$

检验: 当 $\bar{X} \leq \mu_0 - z_\alpha \frac{\sigma}{\sqrt{n}}$ 时拒绝.

Remark. 这也被称为 Z-检验.

e.g. $H_0: \mu \geq 5$, $H_1: \mu < 5$ (i.e. 上侧中 $\mu_0 = 5$). $n = 10$, $\sigma = 0.01$, $\alpha = 0.01$

$$\text{临界值 } c = \mu_0 - z_\alpha \frac{\sigma}{\sqrt{n}} \approx 4.993$$



// 在 N-P 范式中: α 是
// 错误地拒绝
// H 的概率

e.g. 若 σ^2 未知, 则采用 $\frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \sim t(n-1)$

这是 t 检测

Lecture 13

Review:

1 H_0 v.s. H_1

2. 决策: 拒绝或不拒绝 H_0 .

检验即决策准则 $\iff$ 拒绝域 R 的划分

3. 统计学中没有绝对的证实 \ 证伪

$\alpha(R)$ v.s. $\beta(R)$ 是检验程序的属性, 不是样本的属性.

$\alpha(R), \beta(R)$ 都是参数 θ 的函数

$$\alpha(R) = P((X_1, \dots, X_n) \in R \mid H_0)$$

$$\beta(R) = P((X_1, \dots, X_n) \in R^c \mid H_1)$$

α, β 是预先指定的可接受的长期错误率. s.t. $\alpha(R) \leq \alpha, \beta(R) \leq \beta$

4. 功效和功效函数

$$P((X_1, \dots, X_n) \in R) = \begin{cases} \alpha(R) & , \theta \in H_0 \\ \underbrace{1 - \beta(R)}_{\text{功效}} & , \theta \in H_1 \end{cases}$$

$\beta(R)$ 小, 功效大, 则不拒绝 H_0 时, 才能接受 H_1 .

若忽略对 $P(I)$ 的系统控制, 将导致对结果意义及下一步工作的误判

5. 临界值检验法的步骤

Step 1: 提出 H_0, H_1 .

Step 2: 给定 $\alpha > 0$.

Step 3: 确定检验的统计量

确定拒绝域的形状 (看 H_1 , 单边 or 双边)

e.g. $H_0: \theta = \theta_0, H_1: \theta > \theta_0$, 则 R 的形状应为 $\bar{X} \geq c$ 而不是双边情形

Step 4: 由 $P(I) \leq \alpha$ 给出 R

Step 5: 采样, 计算检验统计量的值.

Step 6: 决策

(事实上, 还应该讨论功效)

3° 临界值检验与区间估计的对偶关系.

e.g. $\bar{X} \sim N(\mu, \sigma^2)$, σ^2 已知, $\alpha \in (0, 1)$, $\bar{X}_1, \dots, \bar{X}_n$ iid.

置信区间(双侧) $\mu \in (\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}})$

假设检验: 考虑 $H_0: \mu = \mu_0$, $H_1: \mu \neq \mu_0$.

$$P(|\bar{X} - \mu_0| \geq c) \leq \alpha$$

$\Rightarrow$ 检验: $|\bar{X} - \mu_0| \geq z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ 时拒绝.

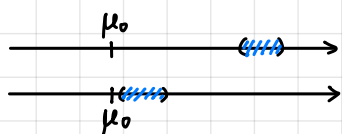
$$\Rightarrow R^c = \{(\bar{X}_1, \dots, \bar{X}_n) \mid |\bar{X} - \mu_0| < z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\}$$

$$\Rightarrow \mu_0 \in (\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}})$$

由此可见: $\mu_0 \in \text{置信区间} \iff \text{假设检验 } (H_0: \mu = \mu_0, H_1: \mu \neq \mu_0) \text{ 不拒绝 } H_0$.

Remark. 区间估计信息更丰富

e.g. 两个拒绝 H_0 的情形, 区间估计给出了更多信息.



4° p值检验法.

e.g. $\bar{X} \sim N(\mu, \sigma^2)$, $\sigma^2 = 25$, 检验假设 $H_0: \mu = 10$, $H_1: (\mu \neq 10)$ (i.e. $\mu_0 = 10$)

$n = 100$, $\bar{X}$ 的观测值 $\bar{x} = 10.935$. 给定 $\alpha \in (0, 1)$

检验 $|\bar{X} - \mu_0| \geq z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ 时拒绝 H_0 .

若取 $\alpha = 0.05$, $|\bar{x} - 10| = 0.935 < z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} = 0.980 \Rightarrow$ 不拒绝 H_0 .

若取 $\alpha = 0.10$, $|\bar{x} - 10| = 0.935 > z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} = 0.825 \Rightarrow$ 拒绝 H_0 .

$$P_{\mu=\mu_0}(|\bar{X} - \mu_0| \geq |\bar{x} - \mu_0|) = P(|Z| \geq \frac{|\bar{x} - \mu_0|}{\frac{\sigma}{\sqrt{n}}})$$

$$\text{即 } P(|Z| \geq 1.87) \approx 0.0614$$

P值: 原假设为真时, 观测值以及更极端的观测出现的概率

检验 $\iff$ p值 $\leq \alpha$ 时拒绝 H_0 . (通常称为观测显著)

Remark. ① p值 $\neq P(H_0 | \bar{x})$

② p值不小 $\Rightarrow$ 不拒绝 H_0 .

原因可能是: a) H_0 确实为真

b) H_0 为假, 但检验功效 ($1 - \beta(LR)$) 太低

P值检验法的步骤

Step 1 提出 H_0, H_1

(Step 2 给定 $\alpha > 0$)

Step 3 确定检验的统计量
确定“极端”的形式. (由 H_1 决定)

Step 4 采样, 计算统计量的值

Step 5 算 p 值

Step 6 决策

e.g. (选举问题, 复合假设的 p 值检验)

调查到的支持率 $\frac{684}{1200} := p_n$ (观测比例), $n=1200$

• $H_0: p = p_0, H_1: p > p_0$ (H_1 决定极端的形状)

当 H_0 为真时: $\frac{p_n - p_0}{\hat{se}(p_n)} \overset{\text{近似}}{\sim} N(0, 1)$, 其中 $\hat{se}(p_n) = \sqrt{\frac{p_n(1-p_n)}{n}}$ 或 $\sqrt{\frac{p_0(1-p_0)}{n}}$

$$p\text{值} = P\left(\frac{p_n - p_0}{\hat{se}(p_n)} \geq \frac{p_n - p_0}{\hat{se}(p_n)}\right) \approx P\left(Z \geq \frac{p_n - p_0}{\hat{se}(p_n)}\right)$$

若 $p_0 = 0.550$, 则 $p\text{值} \approx 0.081$

若 $p_0 = 0.545$, 则 $p\text{值} \approx 0.04$

• $H_0: p \leq p_0, H_1: p > p_1$

当 H_0 为真 ($p \leq p_0$) 时: 更极端形式应为 $p_n \geq p_n$

$$P_{p \leq p_0} \left(\frac{p_n - p_0}{\hat{se}(p_n)} \geq \frac{p_n - p_0}{\hat{se}(p_n)} \right)$$

$$= P_{p \leq p_0} \left(\frac{p_n - p}{\hat{se}(p_n)} \geq \frac{p_n - p_0}{\hat{se}(p_n)} + \frac{p_0 - p}{\hat{se}(p_n)} \right)$$

$$\leq P_{p \leq p_0} \left(Z \geq \frac{p_n - p_0}{\hat{se}(p_n)} \right) \quad (\text{由于 } p_0 - p \geq 0)$$

$$= P_{p = p_0} \left(\frac{p_n - p_0}{\hat{se}(p_n)} \geq \frac{p_n - p_0}{\hat{se}(p_n)} \right) \quad // \text{原概率率在 } p = p_0 \text{ 处取最大值}$$

$$\text{此时: } p\text{值} = \sup_{p \leq p_0} P_{p = p_0} \left(\frac{p_n - p_0}{\hat{se}(p_n)} \geq \frac{p_n - p_0}{\hat{se}(p_n)} \right)$$

• 定理 若拒绝 H_0 , $\theta \in \Theta \iff T(x_1, \dots, x_n) \geq c$
 则 $P\text{值} = \sup_{\theta \in \Theta} P(T(x_1, \dots, x_n) \geq T(x_1, \dots, x_n))$

5° Bayes 假设检验

e.g. 给定硬币, 掷 10 次, 观测到正面向上 x 次

$$H_0: p=0.5, H_1: p=0.7$$

$$\frac{P(H_0|x)}{P(H_1|x)} = \frac{P(H_0)P(x|H_0)}{P(H_1)P(x|H_1)} \stackrel{\text{e.g.}}{<1} \Rightarrow \text{拒绝 } H_0.$$

先验之比 分布之比

Remark. 也可以设置后验概率之比 $< c$ 时拒绝 H_0 . (c 不一定为 1)

Remark. e.g. $H_0: \theta = \theta_0$, Θ 连续, 则 $P(H_0|x) \equiv 0$.

此时需技术处理, 参考陈书 5.2.8, 课程不要求.

6° 拟合优度检验

e.g. 点数	1	2	3	4	5	6	
观测	4	6	17	16	8	9	60

H_0 : 均匀 ($p_i = \frac{1}{6}, \forall i \in \{1, \dots, 6\}$).

H_1 : 不均匀

Pearson: χ^2 统计量: $\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$

O_i : 观测频数
 E_i : 期望频数

定理: $H_0: P(X = a_i) = p_i \quad (i=1, \dots, k)$

若 H_0 为真, 则当 $n \rightarrow \infty$ 时 χ^2 的分布 $\rightarrow \chi^2(k-1)$

本例中: 观测值 $\chi_0^2 = 14.2$

$\Rightarrow P\text{值} = P_{H_0}(\chi^2 \geq \chi_0^2) \approx 0.014$?

Remark. 应用中, 至少要有 $E_i \geq 5$, 才能较好地运用该近似准则

事实上, 是多项分布检验

Lecture 14

Review

1. 假设检验 { Neyman-Pearson < 临界值检验
p值检验 > { $Z, t, \chi^2, F \dots$
拟合优度, 正态性...
似然比, 符号, 秩和...
列联表, ...

Pearson χ^2 检验

e.g. $H_0: p_1: p_2 = 1:1$

$$\chi^2 = \sum_{i=1}^2 \frac{(O_i - E_i)^2}{E_i} = \frac{(O_1 - np_1)^2}{np_1} + \frac{(O_2 - np_2)^2}{np_2} \sim \chi^2(2-1) = \chi^2(1)$$
$$= \frac{(O_1 - np_1)^2}{np(1-p)} \xrightarrow[n \rightarrow \infty]{\text{近似}} \chi^2(1)$$

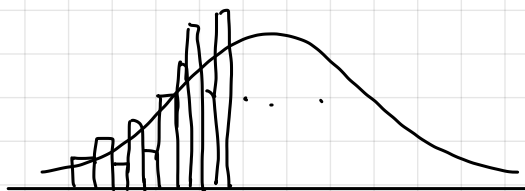
(直观理解)

6° 拟合优度检验

e.g. 汽车电池寿命 $n=40$

.....

分组 $\rightarrow$ 相对频率 $\rightarrow$ 直方图 $\rightarrow$ 总体分布



H_0 : 电池寿命分布是正态的

H_1 : H_0 的否定.

用 Pearson χ^2 检验.

单元边界	O_i	$E_i = np_i$	
1.45 - 1.95	2	0.5	} 8.5 将 E_i 不足 5 的单元进行合并
1.95 - 2.45	1	2.1	
2.45 - 2.95	4	5.9	
...	15	10.3	
...	10	10.7	} 10.5
...	5	7.0	
4.45 ~ 4.95	3	3.5	

基本想法: $P_i(\theta) = \int_{I_i} f(x; \theta) dx$

$$L(\theta) = \prod_{i=1}^k P_i^{O_i}(\theta)$$

→ 求 θ 的极大似然估计

$$\rightarrow P_i \approx P_i(\theta^*) \rightarrow E_i = np_i(\theta^*)$$

$$\chi^2 \sim \chi^2(k-1-s)$$

但 θ 在积分里 θ^* 根本求不出来.

实践中.

① 直接求 $\mu^* = \bar{x}$, $\sigma^* = m_2$.

此例中: $\mu^* = 3.5$, $(\sigma^*)^2 = 0.7^2$

② 计算 E_i

以 E_4 为例: $p_4 = P(2.95 \leq x \leq 3.45)$

$$= P\left(\frac{2.95-3.5}{0.7} \leq \frac{x-3.5}{0.7} \leq \frac{3.45-3.5}{0.7}\right)$$

$$= 0.2573$$

用 MLE 值将 x 标准化

$$E_4 = np_4 = 10.3$$

首尾要特别处理: $p_1 = P(\bar{x} \leq 1.95)$

③ 合并后, $k=4$, $\chi^2_0 = \sum_{i=1}^4 \frac{(O_i - E_i)^2}{E_i} \approx 3.05$

$$\chi^2(k-1-s) = \chi^2(1)$$

取 $\alpha = 0.05$

$$\chi^2_{0.05}(1) = 3.84$$

⇒ 不拒绝

Remark. ① 不拒绝并不意味着一定正态

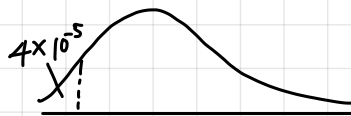
② 上述做法的 χ^2 统计量并不近似为 $\chi^2(k-1-s)$, 估计的参数维数
但检验p值的上界、下界分别是 $\chi^2(k-1-s)$, $\chi^2(k-1)$ 的p值.

独立 χ^2 统计量合并 ~ 自由度相加

e.g. Mendel 的实验全部独立(不同作物)

Fisher 计算每个 χ^2 统计量并合并

得到 χ^2 -统计量略小于42, 自由度为84.



$$P_{\text{值}} = 1 - 4 \times 10^{-5}$$

• 检验数据是否造假

e.g. H_0 : Mendel 的数据是真实的

H_1 : Mendel 的数据是捏造的 过拟合了, χ^2 更小

↓
“极端”的形式: 在本例中, p值越小越极端.

$$P = 4 \times 10^{-5}$$

拒绝 H_0 , 认为 Mendel 数据造假

7° 3 | 联表检验 (contingency-table) 检验

e.g. (独立性检验) 啤酒口味偏好和饮酒者性别是否独立.

样本

	轻淡	普通	浓重	
男	20 ^{26.67}	40	20	80
女	30	30	10	70
	50	70	30	150

a行 × b列

$$H_0: \text{独立: } p_{ij} = p_{i+} \cdot p_{+j}$$

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \quad E_{ij} = n p_{ij}$$

① 极大似然估计 p_{ij}

$$p_{ij}^* = p_{i+}^* p_{+j}^* = \frac{\sum_j O_{ij}}{n} \cdot \frac{\sum_i O_{ij}}{n}$$

$$// \quad p_{i+}^* = \frac{\text{第 } i \text{ 行的和}}{\text{总数 } n}, \quad p_{+j}^* = \frac{\text{第 } j \text{ 列的和}}{\text{总数 } n}$$

$$\textcircled{2} E_{ij} = n p_{ij} \approx n p_{ij}^* = \frac{(\sum_i o_{ij})(\sum_j o_{ij})}{n}$$

观测值 $\chi_0^2 = \sum_{i,j} \frac{(o_{ij} - E_{ij})^2}{E_{ij}} \approx 6.12$

③ 自由度为 $ab - 1 - [(a-1) + (b-1)] = (a-1)(b-1)$

剩下的一个属于自然推断

How?

④ $p\text{值} = P(\chi^2 \geq \chi_0^2) = 0.0469$

e.g. (齐性检验) Jane Austin 与其续写

	Sense & Sensibility	Emma	Sandition I	Σ	Sandition II
a	147	186	101	434	83
an	25	26	11	62	29
this	32	39	15	86	15
that	94	105	37	236	22
with	59	74	28	161	43
without	18	10	10	78	4
	375	440	202	1017	196
	n_1	n_2	n_3		

(1) 检验 Austin 不同作品单词用法的一致性

H_0 : 具有一致性 ($P_{i1} = P_{i2} = P_{i3}$)

在相应总体中 "i" 行对应情形的出现概率

① 在 H_0 真的前提下,

设 $P_{i1} = P_{i2} = P_{i3} = P_i$

对 MLE 估计 $P_i^* = \frac{434}{1017}$ (多项分布概率的 MLE 是当前比例)

$E_{ij} = n_j P_i \approx n_j P_i^*$

e.g. $E_{11} \approx n_1 P_1^* = 375 \times \frac{434}{1017} = 160.0$

$\chi_0^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \approx 12.27$

③ 自由度 $b(a-1) - (a-1) = (a-1)(b-1)$

$\equiv 10$

(?)

$p\text{值} = P(\chi^2 \geq \chi_0^2) \in [0.25, 0.35]$, 不拒绝 H_0

?

(2) 检验仿写者与 Austin 不同单词用法的一致性

$$H_0: P_{ij} = P_i^* \quad (\text{用小计与仿写者检验})$$

$$\text{自由度 } (a-1)(b-1) = 5$$

$$P\text{值} < 10^{-3} \Rightarrow \text{拒绝 } H_0, \text{仿写一致性不行}$$

Lecture 15

8⁰ 似然比检验

e.g. 掷硬币, $n=10$, 正面向上 x 次 $H_0: p=0.5$, $H_1: p=0.7$

考虑 Bayes 检验.

$$\frac{P(H_0|x)}{P(H_1|x)} = \frac{\underbrace{P(H_0)}_{\text{先验比}} \underbrace{P(x|H_0)}_{\text{似然比}}}{P(H_1) P(x|H_1)}$$

似然比 $\frac{P(x|H_0)}{P(x|H_1)} < C$ 时拒绝 H_0

在不考虑先验比的情况下, 用似然比进行检验, 用检验水平 α 决定 C

Remark. 当假设都是简单假设时, Neyman Pearson 证明:

似然比检验是最优的 (即功效最大)

当假设不都是简单假设时,

似然比一般不最优, 但表现不错.

• $H_0: \theta \in \Theta_0$, $H_1: \theta \in \Theta_1$, 样本 i.i.d.

$$\text{广义似然比: } \Lambda^* \triangleq \frac{\sup_{\theta \in \Theta_0} L(\theta)}{\sup_{\theta \in \Theta_1} L(\theta)}$$

基于技术原因, Λ^* 不能直接用, 采用检验统计量

$$\Lambda = \frac{\sup_{\theta \in \Theta_0} L(\theta)}{\sup_{\theta \in \Theta_0 \cup \Theta_1} L(\theta)} = \min \{ \Lambda^*, 1 \}$$

$\sup_{\theta \in \Theta_0 \cup \Theta_1} L(\theta) = L(\theta^*)$, 可套用 MLE 计算

Λ 越小, 则样本越反对 H_0 ("极端"形式, 越小越倾向于 H_1)

• 临界值检验 $\Leftrightarrow$ 选择 λ_0 s.t. $P(\Lambda \leq \lambda_0 | H_0) \leq \alpha$

Remark. 若 H_0 为真时 Λ 分布易求, 则可直接算 $P(\Lambda \leq \lambda_0 | H_0)$
但通常难以计算

定理 在一定条件下 (光滑性), $-2 \ln \Lambda \rightarrow \chi^2$

$$df = \underbrace{\dim(\Theta_1 \cup \Theta_0)}_{\text{自由参数的个数}} - \dim(\Theta_0)$$

自由参数的个数.

e.g. 多项分布 $H_0: P_1 = p_{10}, \dots, P_k = p_{k0}$

各单元观测频数 $n_1, \dots, n_k$

$$L(p_1, \dots, p_k) = \binom{n}{n_1, \dots, n_k} \prod_{i=1}^k p_i^{n_i}$$

$$\Lambda = \frac{L(p_1, \dots, p_k)}{L(p_1^*, \dots, p_k^*)}, \quad p_i^* = \frac{n_i}{n}$$

$$-2 \log \Lambda = -2 \left(\sum_{i=1}^k n_i \ln \frac{p_i}{p_i^*} \right)$$

$$= 2 \left(\sum_{i=1}^k O_i \ln \frac{p_i^* \cdot n}{p_i \cdot n} \right)$$

$$= 2 \left(\sum_{i=1}^k O_i \ln \frac{O_i}{E_i} \right)$$

$$= 2 \sum_{i=1}^k \underbrace{(O_i - E_i)}_{\text{各自求和都为 } n} + \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} + \dots$$

$$= \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$$

$\dim(\Theta_0) = 0$, H_0 各参数已确定, 没有自由度

$\dim(\Theta_0 \cup \Theta_1) = k-1$, 全参数空间, k 参数, 1 约束, $(k-1)$ 自由度

$x \ln \frac{x}{x_0}$ 于 $x = x_0$ 处 Taylor 展开:

$$x \ln \frac{x}{x_0} = 0 + (x - x_0) + \frac{1}{2} \frac{(x - x_0)^2}{x_0} + o((x - x_0)^2)$$

9° 两个总体的比较

• 两个独立总体的比较

总体	均值	方差	样本
X	μ_1	σ_1^2	$X_1, \dots, X_n$
Y	μ_2	σ_2^2	$Y_1, \dots, Y_m$

• 比较均值 $E(\bar{X} - \bar{Y}) = \mu_1 - \mu_2$, $Var(\bar{X} - \bar{Y}) = \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}$

关键在于用 σ_i^2 来估计 $se(\bar{X} - \bar{Y}) = \sqrt{Var(\bar{X} - \bar{Y})}$

① 若 X, Y 都正态, $\sigma_1^2 = \sigma_2^2 = \sigma^2$

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sigma \sqrt{\frac{1}{n} + \frac{1}{m}}} \sim N(0, 1)$$

$$S_1^2 = \frac{1}{n-1} \sum_{i=1}^n (\bar{X} - X_i)^2$$

$$S_2^2 = \frac{1}{m-1} \sum_{i=1}^m (\bar{Y} - Y_i)^2$$

$$S^2 = \frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}$$

$$\text{由 } \frac{(n-1)S_1^2}{\sigma^2} \sim \chi^2(n-1),$$

$$\frac{(m-1)S_2^2}{\sigma^2} \sim \chi^2(m-1).$$

$$\text{有 } \frac{(n+m-2)S^2}{\sigma^2} \sim \chi^2(n+m-2)$$

② 一般地, $\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \xrightarrow[\text{CLT}]{\text{近似}} N(0, 1)$
CLT X, Y 本身无正态性.

e.g. (比较成功率 v.s. 失败率)

Aspirine:

	病发	未发	合计	发病率
安慰剂	239	10795	11034	0.0217 P_1
阿司匹林	139	10898	11037	0.0126 P_2

用假设检验 $H_0: P_1 = P_2$, $H_1: P_1 > P_2$ (Bernoulli 实验 $p \Rightarrow \text{var}$)

$$\frac{(P_1 - P_2) - (p_1 - p_2)}{\widehat{se}(P_1 - P_2)} \xrightarrow[\text{近似}]{\text{理论值}} N(0, 1)$$

统计量 / 其观测值

$$Var(P_1 - P_2) = \frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}$$

$$= [se(P_1 - P_2)]^2$$

H_0 为真时: $P_1 = P_2 = p$

$$\text{Var}(P_1 - P_2) = p(1-p) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$$

$$\text{估计 } p^* = \frac{k_1 + k_2}{n_1 + n_2}$$

↓

$$\text{Var}(P_1 - P_2) = (0.00175)^2$$

$$P\text{值} = P_{H_0} \left(Z > \frac{0.0091 - 0}{0.0175} \approx 5.20 \right) \approx 10^{-4}$$

Remark. ① 随机分布, ② 双盲 / 单盲, ③ n 充分大

• 比较方差

e.g. X, Y 两总体独立, 皆正态, 样本 iid 为 $X_1, \dots, X_n; Y_1, \dots, Y_m$.

假设检验: $H_0: \sigma_1^2 = \sigma_2^2, H_1: \sigma_1^2 \neq \sigma_2^2$.

$$\begin{aligned} \frac{(n-1)S^2}{\sigma_1^2} &\sim \chi^2(n-1) \\ \frac{(m-1)S^2}{\sigma_2^2} &\sim \chi^2(m-1) \end{aligned} \quad \text{独立.}$$

$$\text{检验统计量 } F \triangleq \frac{S_1^2}{S_2^2}$$

当 H_0 为真时, $F \sim F(n-1, m-1)$
可约去 σ_1^2, σ_2^2

检验: $F > F_{\frac{\alpha}{2}}(n-1, m-1)$ 或 $F < F_{1-\frac{\alpha}{2}}(n-1, m-1)$

注: ① F 越大,

$$\text{② } F_{\frac{\alpha}{2}}(n-1, m-1) = \frac{1}{F_{1-\frac{\alpha}{2}}(m-1, n-1)}$$

交换上下分位数.

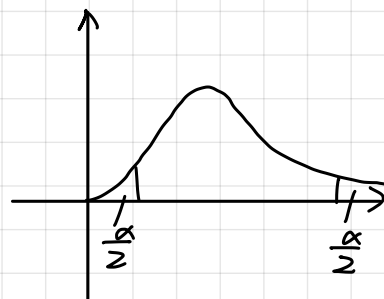
↓
交换自由度再取倒数

设 $W_1 \sim \chi^2(k_1)$

$W_2 \sim \chi^2(k_2)$

W_1, W_2 独立.

$$\frac{\frac{W_1}{k_1}}{\frac{W_2}{k_2}} \sim F(k_1, k_2)$$



e.g. 某大型出租车公司比较汽油A和B的行驶里程

将100辆车均分为2组

	样本容量	平均里程	样本标准差
油A	50	25	5.00
油B	50	26	4.00

→ 标准差较大

差异到?

无法联系, 赶紧

$$H_0: \mu_A = \mu_B \quad H_1: \mu_A \neq \mu_B$$

$$P\text{值} = P(|Z| \geq \left| \frac{25-26-0}{0.905} \right|)$$

$$\approx P(|Z| \geq 1.1) = 0.1357 \times 2 = 0.2714$$

$p > 0.05$, 不拒绝

$\Rightarrow$ 有利于任何一种汽油的证据都不充分

$$E(\bar{X} - \bar{Y}) = \mu_A - \mu_B$$

$$\begin{aligned} \text{Var}(\bar{X} - \bar{Y}) &= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \\ &\approx \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} \\ &\approx (0.905)^2 \end{aligned}$$

改进: 同一司机, 不同日期分配不同的油

$n=10$ 就够了, 也排除了车况/驾驶习惯的影响.

车号	油A	油B	差异
1	27.01	26.95	0.06
2	20.00	20.44	-0.44
...	...	...	...
10	25.22	26.01	-0.79
均值	25.00	25.80	-0.60
标准差	4.27	4.00	0.61

两组相关数据

用差异列检验

差异列标准差变小

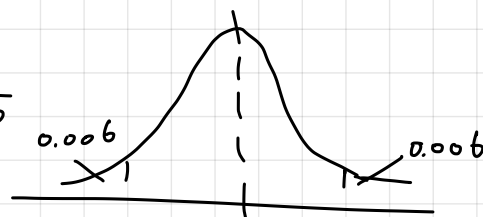
但此时, 油A, 油B两列相关

$$H_0: \mu_d = 0, \mu_d \neq 0, E(\bar{d}) = E(\bar{X} - \bar{Y}) = \mu_A - \mu_B = \mu_d$$

$$\bar{d} = \frac{1}{n} \sum_{i=1}^n d_i, \text{ 假设 } d = X - Y \text{ 服从正态分布.}$$

$$\frac{\bar{d} - \mu_d}{\frac{S_d}{\sqrt{n}}} \sim t(n-1)$$

$$\text{设 } H_0 \text{ 为真, 观测值 } \frac{-0.60 - 0}{\frac{0.61}{\sqrt{10}}} = -3.15$$



$$P\text{值} \approx 0.006 \times 2 = 0.012 < \alpha = 0.05$$

拒绝 H_0

Remark. 假设检验不能检验试验设计

10° 对显著性的思考

(1) 统计显著 $\neq$ 实际显著: 小差异架不住样本多

(2) 数据窥探: 先看数据后定准则

(3) 检验不能检验原因

未尽话题

1. 抽样理论
2. 随机过程
3. 随机模拟 $\begin{cases} MC \\ Bootstrap \end{cases}$
4. 线性模型
5. 方差分析与实验设计 (Fisher 的作品)
6. 假设检验的深入理论
7. 非参数统计

荐书: 1. Larry Wasserman: ① 统计学完全教程

② 现代非参统计

2. B. Efron, T. Hastie: ① 计算机时代的统计推断