数理统计

第 4 章　参数估计

4.1 基本概念与性质

4.1.1 基本概念与常用统计图

总体（母体）是概率分布族的一员.

总体分布 离散性（概率函数），连续型（概率密度函数）

单参数分布族

非参数总体

样本大小 (容量)

常用统计图

1. **频数分布表**

组号	区间	频数 n_i	频率 f_i
1	$(1,2]$	2	0.40
2	$(2,3]$	3	0.60
合计		5	1.00

2. **频率直方图** 以 $\dfrac{f_i}{\Delta t_i}$ 为高. 所有小矩形的面积和为 1.
3. **条形图**　一般用于小样本离散性随机变量总体分布.

4.1.2 经验分布函数与格列文科定理

经验分布函数 (样本分布函数) $F_n(x)=\{X_1,X_2,\cdots,X_n \text{ 中不大于 } x \text{ 的个数}\}/n.$

即将 X 的样本值 $x_1,x_2,\cdots,x_n$ 从小到大重排后, 定义经验分布函数如下.

$$\underbrace{x_{(1)},\cdots,x_{(1)}}_{n_1},\underbrace{x_{(2)},\cdots,x_{(2)}}_{n_2},\cdots\underbrace{x_{(m)},\cdots,x_{(m)}}_{n_m},$$

$$F_n(x)=\begin{cases}0, & x<x_{(1)},\\ \dfrac{1}{n}\displaystyle\sum_{i=1}^{k}n_i, & x_{(k)}\le x<x_{(k+1)}\\ 1, & x\ge x_{(m)}\end{cases}$$

1. $0\le F_n(x)\le 1.$
2. $F_n(x)$ 单调不减.
3. $F_n(-\infty)=0,\ F_n(+\infty)=1.$

4. $F_n(x)$ 右连续.

格列文科定理 对于任意实数 x, 经验分布函数 $F_n(x)$ 以概率 1 一致收敛于总体分布函数 $F(x)$, 即

$$P\left\{\lim_{n\to\infty}\sup_{-\infty<x<+\infty}|F_n(x)-F(x)|=0\right\}=1.$$

- 经验分布函数不确定, 不唯一, 所以在极限外套一个 P.

4.1.3　常用统计量及其性质

统计量　只依赖于样本，而不依赖于其未知参数.

- 样本的统计量为 $g(X_1,X_2,\cdots,X_n)$.
- 统计量的观测值为 $g(x_1,x_2,\cdots,x_n)$.

样本均值 $\overline{X}=a_1=\dfrac{1}{n}\sum\limits_{i=1}^{n}X_i.$

- 其观测值记为 $\overline{x}=\dfrac{1}{n}\sum\limits_{i=1}^{n}x_i.$
- 对于一般分布 ⭐
  - $E(\overline{X})=\mu.$
  - $E\left(\overline{X}^2\right)=\dfrac{\sigma^2}{n}+\mu^2.$
  - $E\left(\overline{X}^3\right)=\dfrac{\mu_3}{n^2}+\dfrac{3}{n}\mu\sigma^2+\mu^3.$
  - $E\left(\overline{X}^4\right)=\dfrac{\mu_4}{n^3}+\dfrac{4}{n^2}\mu_3\mu+\dfrac{18}{n}\mu^2\sigma^2+7\mu^4.$
  - $E\left(X_i^2\overline{X}\right)=\dfrac{\mu_3}{n}+\dfrac{n-1}{n}(\sigma^2+\mu^2)\mu.$
  - $E\left(X_i\overline{X}^{k-1}\right)=E\left(\overline{X}^k\right).$
  - $D(\overline{X})=\dfrac{\sigma^2}{n}.$
  - $\mu_k(\overline{X})=E\left[\sum\limits_{i=1}^{n}\left(X_i-\overline{X}\right)^k\right]=\dfrac{\mu_k}{n^{k-1}}.$
- ⭐ 设总体有 N 个数据 $a_1,a_2,\cdots,a_N$, 均值为 $\mu=\dfrac{1}{N}\sum\limits_{i=1}^{N}a_i$, 方差为 $\sigma^2=\dfrac{1}{N}\sum\limits_{i=1}^{N}(a_i-\mu)^2.$

  从总体中抽取 n 个值 $X_1,X_2,\cdots,X_n$ 作为样本, 则
  - $E(\overline{X})=\mu.$
  - $D(\overline{X})=\dfrac{N-n}{N-1}\dfrac{\sigma^2}{n}.$

  当 $N\to+\infty$ 时, 即抽取的样本相互独立时, 有 $D(\overline{X})\sim\dfrac{\sigma^2}{n}.$

样本离差平方和 $\mathrm{SS}=\sum_{i=1}^{n}(X_i-\overline{X})^2=\sum_{i=1}^{n}X_i^2-n\overline{X}^2$.

- 其观测值记为 $\mathrm{ss}=\sum_{i=1}^{n}(x_i-\overline{x})^2=\sum_{i=1}^{n}x_i^2-n\overline{x}^2$.

样本方差 $S^2=\dfrac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2$.

- 其观测值记为 s^2.
- 标准差又称均方差, 样本标准差的观测值记为 s.
- 对于一般分布, 有
  - $E(S^2)=\sigma^2$.
  - $D(S^2)=\dfrac{1}{n}\left(\mu_4-\dfrac{n-3}{n-1}\sigma^4\right)$. ⭐
  - $\mathrm{Cov}(\overline{X},S^2)=\dfrac{\mu_3}{n}$. ⭐
- 对于正态分布, 有
  - $D(S^2)=D\left(\dfrac{\sigma^2}{n-1}\chi_{n-1}^2\right)=\dfrac{2\sigma^4}{n-1}$. ⭐
  - $\mathrm{Cov}(\overline{X},S^2)=0$.

样本原点矩 $a_k=\dfrac{1}{n}\sum_{i=1}^{n}X_i^k$.

- 或使用 A_k 表示样本原点矩, 用 a_k 表示其观测值.

样本中心矩 $m_k=\dfrac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^k$.

- 或使用 B_k 表示样本中心距, 用 b_k 表示其观测值.
- 矩称为理论矩，样本矩称为经验矩，即经验分布函数的矩.
- $m_2=\dfrac{n-1}{n}S^2$.
- $E(m_3)=\left(1-\dfrac{4}{n}+\dfrac{2}{n^2}\right)\mu_3+\dfrac{9-3n}{n}\mu\sigma^2+\dfrac{3-n}{n}\mu^3$.

不知道叫什么的统计量 $\sum_{i=1}^{n}(X_i-\mu)^2=\mathrm{SS}+n(\overline{X}-\mu)^2$.

样本协方差 $S_{XY}=\dfrac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})$.

- 其观测值记为 $s_{_{XY}}$.
- 样本协方差是总体协方差的无偏估计 (可以是有限总体的不放回抽取)

  $E(S_{XY})=\mathrm{Cov}(X,Y)$.

样本相关系数 $\rho_{XY}=\dfrac{S_{XY}}{S_XS_Y}$.

- 其中 S_X 和 S_Y 为样本均方差. 其观测值为 $\rho_{XY} = \dfrac{s_{XY}}{s_X s_Y}$.

样本的 **众数** 记为 M_0.

次序统计量 $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$.

样本中位数 $\hat{m} = M_e = \begin{cases} X_{((n+1)/2)}, & n \text{ 为奇数}, \\ (X_{(n/2)} + X_{(n/2+1)})/2, & n \text{ 为偶数}. \end{cases}$

偏态系数

- 计算公式
  - 简单偏态系数 $\mathrm{SK} = \dfrac{m_3}{\sigma^3} = \dfrac{\sum(X-\overline{X})^3}{\sigma^3 \cdot N}$.
  - 加权偏态系数 $\mathrm{SK} = \dfrac{\sum(X-\overline{X})^3 F}{\sigma^3 \sum F}$.
- 取值说明
  - $\mathrm{SK} = 0$ 表示数据为完全的对称分布.
  - $\mathrm{SK} > 0$ 表示数据为 **正偏态** (或 **右偏态**).
  - $\mathrm{SK} < 0$ 表示数据为 **负偏态** (或 **左偏态**).

4.3.4 正态总体常用统计量

设 $X_1, X_2, \cdots, X_n$ 与 $Y_1, Y_2, \cdots, Y_n$ 分别是来自正态总体 $N(\mu_1, \sigma_1^2) = N(\mu, \sigma^2)$ 和 $N(\mu_2, \sigma_2^2)$ 的相互独立的两个样本, 则

- 样本均值
  - $\overline{X} \sim N\left(\mu, \dfrac{\sigma^2}{n}\right)$.
  - $\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$.
  - $\displaystyle\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 \sim \chi_n^2$.
- 样本方差
  - $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$.
  - $\overline{X}$ 与 S^2 相互独立.
  - $\dfrac{\overline{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$.
- 两组样本
  - $\overline{X} - \overline{Y} \sim N\left(\mu_1 - \mu_2, \dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}\right)$.

- $\dfrac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \sim f_{n_1-1,n_2-1}$.
- 当 $\sigma_1^2=\sigma_2^2=\sigma^2$, 令 $S_\omega^2=\dfrac{\mathrm{SS}_1^2+\mathrm{SS}_2^2}{n_1+n_2-2}$, 则

$$\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{S_\omega\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}} \sim t_{n_1+n_2-2}.$$

⭐ 设有 m 个方差均为 σ^2 的分布, 其中有 k 个分布的期望未知. 从这些分布中任取 n 个相互独立的样本数据, 则样本方差的一个无偏估计是 $S^2=\dfrac{\mathrm{SS}}{n-k}$. 若这些分布都是正态分布, 则进一步有 $\dfrac{\mathrm{SS}}{\sigma^2}\sim\chi^2_{n-k}$.

4.2 点估计

4.2.1 矩估计法

$$\alpha_m=\int_{-\infty}^{+\infty}x^m f(x;\theta_1,\cdots,\theta_2)\,\mathrm{d}x\approx a_m=\sum_{i=1}^n\frac{X_i^m}{n}.$$

取 $m=1,2,\cdots,k$，联立方程组即得 $\theta_i\approx\hat{\theta}_i(X_1,\cdots,X_n)$.

- $\hat{\theta}_i(X_1,X_2,\cdots,X_n)$ 称为矩估计量.
- $\hat{\theta}_i(x_1,x_2,\cdots,x_n)$ 称为矩估计值. (其它估计相关名称类似)

变异系数 σ/μ.

🌙 对于任意均值 μ 与方差 σ^2 存在的总体, 有矩估计:

$$\begin{cases}\mu=\dfrac{1}{n}\displaystyle\sum_{i=1}^n X_i\\ \sigma^2+\mu^2=\dfrac{1}{n}\displaystyle\sum_{i=1}^n X_i^2\end{cases}\quad\Rightarrow\quad\begin{cases}\hat{\mu}=\overline{X}\\ \hat{\sigma^2}=\dfrac{\mathrm{SS}}{n}\end{cases}$$

4.2.2 极大似然估计法

样本 $(X_1,X_2,\cdots,X_n)$ 的总体分布函数 (样本似然函数) 为

$$\begin{aligned}L(x_1,\cdots,x_n,;\theta_1,\cdots,\theta_k)&=\prod_{i=1}^n f(x_i;\theta_1,\theta_2,\cdots,\theta_n)\\ L(X_1,\cdots,X_n;\theta_1^*,\cdots,\theta_k^*)&=\max_{\theta_1,\cdots,\theta_k}L(X_1,\cdots,X_n;\theta_1,\cdots,\theta_k)\end{aligned}$$

欲得到极大似然估计，解如下似然方程组

$$\frac{\partial\ln L}{\partial\theta_i}=0\quad(i=1,2,\cdots,k)$$

4.2.3 贝叶斯估计法

先验分布，先验概率. 允许使用主观概率.

设总体有概率密度 $f(X,\theta)$，抽样本 $X_1,X_2,\cdots,X_n$，则 $(\theta,X_1,\cdots,X_n)$ 的联合密度为

$$h(\theta)f(X_1,\theta)\cdots f(X_n,\theta)$$

由此算出 $(X_1,X_2,\cdots,X_n)$ 的边缘密度为

$$p(X_1,X_2,\cdots,X_n)=\int h(\theta)f(X_1,\theta)\cdots f(X_n,\theta)\mathrm{d}\theta$$

从而得出 θ 在给定 $X_1,X_2,\cdots,X_n$ 的条件密度为

$$h(\theta\mid X_1,\cdots,X_n)=h(\theta)f(X_1,\theta)\cdots f(X_n,\theta)/p(X_1,\cdots,X_n)$$

一般取上式的均值作为估计, 即

$$\tilde{\theta}=E(h(\theta\mid X_1,X_2,\cdots,X_n)).$$

$h(\theta)$ 一般是概率函数，即满足 $h(\theta)\geq 0,\int h(\theta)\mathrm{d}\theta=1.$

但对于积分域为无穷区间，或一些特定的分布，可以采用其它函数，比如 $h(\theta)=1$，或直接取为先验密度等. 此时 $h(\theta)$ 称为 "广义先验密度".

根据 n 次独立试验中事件 A 发生的次数 X 去估计其发生的概率 p，按照 "同等无知" 原则（贝叶斯原则），由上述方法积分得

$$\tilde{p}=\frac{X+1}{n+2}.$$

- 估计正态分布 $N(\mu,\sigma^2)$ 中的 μ 时，取 $h(\mu)=1$;
- 估计正态分布 $N(\mu,\sigma^2)$ 中的 σ 时，取 $h(\sigma)=\sigma^{-1}$;
- 估计指数分布 $E(\lambda)$ 中的 λ 时，取$h(\lambda)=\lambda^{-1}$.

由先验分布 $N(\mu.\sigma^2)$ 估计正态总体 $N(\theta,1)$ 中的 θ 为（取 $h(\theta)$ 为先验密度）

$$\tilde{\theta}=\frac{n}{n+\sigma^{-2}}\overline{X}+\frac{\sigma^{-2}}{n+\sigma^{-2}}\mu.$$

4.3 点估计的优良性准则

1. 小样本性质
    1. 无偏性: 误差期望为零.
    2. 渐进无偏性: 误差期望依概率收敛至零.
2. 有效性 (数量指标)

    1. 最小均方误差估计
    2. 最小方差无偏估计 (MVU)
3. 大样本性质
    1. 相合性 (一致性): 误差依概率收敛至零.
    2. 渐进正态性: 自身分布趋于正态分布

4.3.1 估计量的无偏性

无偏性即无系统误差 $E(\hat\theta)-\theta$. 故 **无偏估计量** $\hat g$ 须满足

$$E_{\theta_1,\cdots,\theta_k}[\hat g(X_1,\cdots,X_n)]=g(\theta_1,\cdots,\theta_k).$$

- $m=\overline{X}$ 是 $E(X)$ 的无偏估计.
- 如果总体均值未知, 则 $S^2=\displaystyle\sum_{i=1}^{n}\frac{(X_i-\overline{X})^2}{n-1}$ 是 $\operatorname{Var}(X)$ 的无偏估计.
- 如果总体均值已知, 则 $\displaystyle\sum_{n=1}^{\infty}\frac{(X_i-\mu)^2}{n}$ 是 $\operatorname{Var}(X)$ 的无偏估计.
- 由 $\sigma^2=E(S^2)=\operatorname{Var}(S)+(ES)^2$ 知, S 总是 σ 系统性偏低的估计. 🌙

设 $\hat g=\hat g(X_1,X_2,\cdots,X_n)$ 是未知参数的函数 $g(\theta_1,\theta_2,\cdots,\theta_k)$ 的一个估计量, 如果 $E(\hat\theta)$ 存在, 且

$$\lim_{n\to\infty}E(\hat g(X_1,X_2,\cdots,X_n))=g(\theta_1,\theta_2,\cdots,\theta_k)$$

则称 $\hat g$ 为 g 的 **渐进无偏估计量**.

- $m_2=\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(X_i-\overline{X})^2$ 是 σ^2 的渐进无偏估计量, 但不是无偏估计量.

- ⭐ 对于正态分布总体 $N(\mu,\sigma^2)$, 由 $(n-1)S^2/\sigma^2\sim\chi^2_{n-1}$ 算出

$$S/\sigma\sim g(s)=\begin{cases}\dfrac{(n-1)^{\frac{n-1}{2}}}{2^{\frac{n-3}{2}}\Gamma\left(\dfrac{n-1}{2}\right)}\mathrm{e}^{-\frac{n-1}{2}s^2}s^{n-2}, & s>0\\ 0 & s\le 0.\end{cases}$$

计算 $E(S)=\sigma\displaystyle\int_0^{+\infty}sg(s)\,\mathrm{d}s$, 故 σ 的一个无偏估计是

$$\tilde\sigma=\sqrt{\frac{n-1}{2}}\frac{\Gamma\left(\dfrac{n-1}{2}\right)}{\Gamma\left(\dfrac{n}{2}\right)}S.$$

- 无偏估计不一定好, 比如 $X\sim P(\lambda)$, 则 $g(\lambda)=\mathrm{e}^{-2\lambda}$ 唯一的无偏估计为

$$\hat g(X)=\begin{cases}1, & X\text{ 为偶数},\\ -1, & X\text{ 为奇数}.\end{cases}$$

4.3.2 最小方差无偏估计

1 最小均方误差估计

$$
\begin{aligned}
M_{\hat{\theta}}(\theta) &= E_\theta\left[\hat{\theta}(X_1,\cdots,X_n)-\theta\right]^2 \\
&= \mathrm{Var}_\theta(\hat{\theta}) + \left[E_\theta(\hat{\theta})-\theta\right]^2
\end{aligned}
$$

当 $\hat{\theta}$ 是无偏估计量时, $M_{\hat{\theta}}(\theta)=\mathrm{Var}_\theta(\hat{\theta})$.

2 最小方差无偏估计 (MVU 估计)

Minimum Variance Unbiased.

无偏估计中的最小方差，其均方误差不一定最小.

3 克拉美 - 劳不等式

对于单参数情况 $f(x,\theta)$, 为估计 $g(\theta)$, 记 **费歇尔信息量** 为

$$
\begin{aligned}
I(\theta) &= E\left[\left(\frac{\partial f(x,\theta)}{\partial\theta}\Big/ f(x,\theta)\right)^2\right] \\
&= \int\left[\left(\frac{\partial f(x,\theta)}{\partial\theta}\right)^2\Big/ f(x,\theta)\right]\mathrm{d}x \qquad (\text{连续的总体分布}) \\
&= \sum_i\left(\frac{\partial f(a_i,\theta)}{\partial\theta}\right)^2\Big/ f(a_i,\theta) \qquad (\text{离散的总体分布})
\end{aligned}
$$

则对任一无偏估计 $\hat{g}=\hat{g}(X_1,X_2,\cdots,X_n)$, 有 **克拉美 - 劳不等式**:

$$
\mathrm{Var}_\theta(\hat{\theta}) \geq \frac{g'(\theta)^2}{nI(\theta)}.
$$

- MVU 的均方误差不一定是最小的, 如对于正态分布 $N(\mu,\sigma^2)$, 其中 μ 已知, 则 S^2 是 MVU 估计, 但 $E(S_2-\sigma^2)^2=\dfrac{2\sigma^4}{n} > E(\dfrac{n-1}{n+1}S^2-\sigma^2)^2=\dfrac{2\sigma^4}{n+1}$.
- 若 $\hat{\theta}_1$ 和 $\hat{\theta}_2$ 都是 θ 的 MVU 估计, 则 $a\hat{\theta}_1+b\hat{\theta}_2+c$ 是 $(a+b)\theta+c$ 的 MVU 估计. ⭐ ⭐ (利用第三章定理 3.1, 2°)
- 若 $E(X)=\theta$, $\sum_{i=1}^n c_i=1$, 则 $\sum_{i=1}^n c_iX_i$ 是 θ 的无偏估计, 并且当且仅当 $c_i=\dfrac{1}{n}$ 时, 其为 MVU 估计.

4.3.3 估计量的相合性与渐进正态性

1 相合性

如果当样本大小 n 无限增加时, 估计量 $T(X_1,X_2,\cdots,X_n)$ 依概率收敛于被估计值, 则称该估计量是 **相合估计量**, 即

$$
\forall\varepsilon>0:\lim_{n\to\infty}P_{\theta_1,\cdots,\theta_k}\left(\left|\hat{g}(X_1,\cdots,X_n)-g(\theta_1,\cdots,\theta_n)\right|\geq\varepsilon\right)=0.
$$

具有相合性的例子

- 所有的矩估计量.
- 绝大多数极大似然估计.

对于 g 的无偏估计 $\hat{g}$, 由切比雪夫不等式知:

$$0 \le \lim_{n\to\infty} P(|\hat{g}-g| \ge \varepsilon) \le \frac{D(\hat{g})}{\varepsilon^2}.$$

若 $\lim_{n\to\infty} D(\hat{g}) = 0$, 则其为相合估计量. ⭐

2 渐进正态性

参考这篇高等数理统计的文章.

4.4 区间估计

4.4.1 基本概念

奈曼理论 的原则: 先保证可靠度, 再提升精度.

称区间估计 $[\hat{\theta}_1, \hat{\theta}_2]$ 的 **置信系数** 为 $1-\alpha$, 如果

$$\exists \alpha > 0,\ \forall \theta : P_\theta\left(\hat{\theta}_1(X_1,\cdots,X_n) \le \theta \le \hat{\theta}_2(X_1,\cdots,X_n)\right) = 1-\alpha.$$

称区间估计 $[\hat{\theta}_1, \hat{\theta}_2]$ 的 **置信水平** (**置信度**) 为 $1-\alpha$, 如果

$$\exists \alpha > 0,\ \forall \theta : P_\theta\left(\hat{\theta}_1(X_1,\cdots,X_n) \le \theta \le \hat{\theta}_2(X_1,\cdots,X_n)\right) \ge 1-\alpha.$$

- 随机区间 $[\hat{\theta}_1, \hat{\theta}_2]$ 称为 **双侧置信区间**.
- $(-\infty, \hat{\theta}_2]$ 和 $[\hat{\theta}_1, +\infty)$ 称为 **单侧置信区间**.
- $\hat{\theta}_1$ 和 $\hat{\theta}_2$ 分别称为置信下界和置信上界.
- α 一般取为 0.1, 0.05, 0.025, 0.01, 0.001.

区间估计的研究对象:

1. 置信系数或置信水平.
2. 区间长度.
3. 区间右端点与左端点之比.

4.4.2 枢轴变量法

上 β 分位点: $F(v_\beta) = 1-\beta$.

下 β 分位点: $F(w_\beta) = \beta$.

上 β 分位点就是下 $1-\beta$ 分位点.

统计三大分布的上 β 分位点记为: $\chi_n^2(\beta),\ t_n(\beta),\ f_{n,m}(\beta)$.

利用上 β 分位点 w_β 寻找区间估计的 **枢轴变量法**:

1. 找一个与被估计参数 $g(\theta)$ 有关的统计量 T.
2. 找 **枢轴变量** $S(T,g(\theta))$, 使其分布 F 与 θ 无关.
3. $a \le S(T,g(\theta)) \le b \quad \Leftrightarrow \quad A \le g(\theta) \le B$.
4. $P(w_{1-\alpha/2} \le S(T,g(\theta)) \le w_{a/2}) = 1-\alpha$.

一样本 t 区间估计, 为保证长度 $2St_{n-1}(\alpha/2)/\sqrt{n} \le L$, 斯泰因提出了 **两阶段抽样** 的方法, 其中追加抽样的次数为

$$m = \begin{cases} 0, & n \le [4t_{n-1}^2(\alpha/2)S^2/L^2]+1, \\ n-1-[4t_{n-1}^2(\alpha/2)S^2/L^2], & n > [4t_{n-1}^2(\alpha/2)S^2/L^2]+1. \end{cases}$$

记两次样本全体的均值为 $\tilde{X}$, 则区间估计 $[\tilde{X}-L/2,\ \tilde{X}+L/2]$ 有置信系数 $1-\alpha$.

4.4.3 大样本法

大样本区间估计: 利用 中心极限定理 与枢轴变量法.

例如, 一般的, 设总体有均值 θ, 方差 σ^2, 并且都位置, 从样本 $X_1, X_2, \cdots, X_n$ 做 θ 的区间估计. 由于样本均方差 S 是 σ 的祥和估计, 利用中心极限定理, 当 n 足够大时, 有

$$\sqrt{n}(\overline{X}-\theta)/S \sim N(0,1).$$

以此为枢轴变量, 于是有区间估计

$$\left[\overline{X}-Su_{\alpha/2}/\sqrt{n},\ \overline{X}+Su_{\alpha/2}/\sqrt{n}\right].$$

对于二项分布, 当 $\alpha=0.05,\ n \ge 40$ 时, 有区间长度 $\theta_2-\theta_1 \le 0.3$.

4.4.4 置信界

置信系数 (水平) 为 α 的置信上界 $\overline{\theta}$ 和置信下界 $\underline{\theta}$:

$$\begin{aligned} \forall\theta : P_\theta(\overline{\theta}(X_1,X_2,\cdots,X_n) \ge \theta) = 1-\alpha \\ \forall\theta : P_\theta(\underline{\theta}(X_1,X_2,\cdots,X_n) \le \theta) = 1-\alpha \end{aligned}$$

4.4.5 贝叶斯法

即寻找 $\hat{\theta}_1,\ \hat{\theta}_2$, 使得

$$\int_{\hat{\theta}_1}^{\hat{\theta}_2} h(\theta \mid X_1, \cdots, X_n)\, \mathrm{d}\theta = 1 - \alpha \qquad \text{(区间估计)}$$

$$\int_{-\infty}^{\hat{\theta}} h(\theta \mid X_1, \cdots, X_n)\, \mathrm{d}\theta = 1 - \alpha \qquad \text{(置信上界)}$$

$$\int_{\hat{\theta}}^{+\infty} h(\theta \mid X_1, \cdots, X_n)\, \mathrm{d}\theta = 1 - \alpha \qquad \text{(置信下界)}$$

区间估计中确定 θ_1, θ_2 的方法 (原则):

1. 使 $\hat{\theta}_2 - \hat{\theta}_1$ 最小.
2. 使 $\hat{\theta}_2 / \hat{\theta}_1$ 最小.
3. 取置信水平为 $\alpha/2$ 的置信上下界.

常用区间估计表

表 7.1 正态总体均值、方差的置信区间(置信度为 $1-\alpha$)

	待估参数	其他参数	置信函数 G 的分布	双侧置信区间	单侧置信区间
单个正态总体	μ	σ^2 已知	$\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$	$\left(\overline{X}-\dfrac{\sigma}{\sqrt{n}}z_{\alpha/2}, \overline{X}+\dfrac{\sigma}{\sqrt{n}}z_{\alpha/2}\right)$	$\left(-\infty, \overline{X}+\dfrac{\sigma}{\sqrt{n}}z_{\alpha}\right), \left(\overline{X}-\dfrac{\sigma}{\sqrt{n}}z_{\alpha}, +\infty\right)$
		σ^2 未知	$\dfrac{\overline{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$	$\left(\overline{X}-\dfrac{S}{\sqrt{n}}t_{\alpha/2}(n-1), \overline{X}+\dfrac{S}{\sqrt{n}}t_{\alpha/2}(n-1)\right)$	$\left(\overline{X}-\dfrac{S}{\sqrt{n}}t_{\alpha}(n-1), +\infty\right), \left(-\infty, \overline{X}+\dfrac{S}{\sqrt{n}}t_{\alpha}(n-1)\right)$
	σ^2	μ 已知	$\dfrac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2 \sim \chi^2(n)$	$\left(\dfrac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{\alpha/2}(n)}, \dfrac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{1-\alpha/2}(n)}\right)$	$\left(\dfrac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{\alpha}(n)}, +\infty\right), \left(-\infty, \dfrac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{1-\alpha}(n)}\right)$
		μ 未知	$\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$	$\left(\dfrac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)}, \dfrac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)}\right)$	$\left(\dfrac{(n-1)S^2}{\chi^2_{\alpha}(n-1)}, +\infty\right), \left(0, \dfrac{(n-1)S^2}{\chi^2_{1-\alpha}(n-1)}\right)$
两个正态总体	$\mu_1-\mu_2$	σ_1^2, σ_2^2 已知	$\dfrac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}} \sim N(0,1)$	$\left(\overline{X}-\overline{Y} \mp z_{\alpha/2}\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}\right)$	$\left(\overline{X}-\overline{Y}-z_{\alpha}\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}, +\infty\right)$, $\left(0, \overline{X}-\overline{Y}+z_{\alpha}\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}\right)$
		$\sigma_1^2=\sigma_2^2=\sigma^2$ 未知	$\dfrac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{S_{\omega}\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}} \sim t(n_1+n_2-2)$	$\left(\overline{X}-\overline{Y} \mp t_{\alpha/2}(n_1+n_2-2)S_{\omega}\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}\right)$	$\left(\overline{X}-\overline{Y}-t_{\alpha}(n_1+n_2-2)S_{\omega}\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}, +\infty\right)$, $\left(-\infty, \overline{X}-\overline{Y}+t_{\alpha}(n_1+n_2-2)S_{\omega}\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}\right)$
	$\dfrac{\sigma_1^2}{\sigma_2^2}$	μ_1, μ_2 已知	$\dfrac{n_1\sigma_1^2}{n_2\sigma_2^2} \cdot \dfrac{\sum_{i=1}^{n_2}(Y_i-\mu_2)^2}{\sum_{i=1}^{n_1}(X_i-\mu_1)^2} \sim F(n_2, n_1)$	$\left(\dfrac{n_2}{n_1} \cdot \dfrac{\sum_{i=1}^{n_1}(X_i-\mu_1)^2}{\sum_{i=1}^{n_2}(Y_i-\mu_2)^2}F_{1-\alpha/2}(n_2, n_1), \dfrac{n_2}{n_1} \cdot \dfrac{\sum_{i=1}^{n_1}(X_i-\mu_1)^2}{\sum_{i=1}^{n_2}(Y_i-\mu_2)^2}F_{\alpha/2}(n_2, n_1)\right)$	$\left(\dfrac{n_2}{n_1} \cdot \dfrac{\sum_{i=1}^{n_1}(X_i-\mu_1)^2}{\sum_{i=1}^{n_2}(Y_i-\mu_2)^2}F_{1-\alpha}(n_2, n_1), +\infty\right)$, $\left(0, \dfrac{n_2}{n_1} \cdot \dfrac{\sum_{i=1}^{n_1}(X_i-\mu_1)^2}{\sum_{i=1}^{n_2}(Y_i-\mu_2)^2}F_{\alpha}(n_2, n_1)\right)$
		μ_1, μ_2 未知	$\dfrac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \sim F(n_1-1, n_2-1)$	$\left(\dfrac{S_1^2}{S_2^2}F_{1-\alpha/2}(n_2-1, n_1-1), \dfrac{S_1^2}{S_2^2}F_{\alpha/2}(n_2-1, n_1-1)\right)$	$\left(\dfrac{S_1^2}{S_2^2}F_{1-\alpha}(n_2-1, n_1-1), +\infty\right)$, $\left(0, \dfrac{S_1^2}{S_2^2}F_{\alpha}(n_2-1, n_1-1)\right)$

表 7.2 0-1 分布参数 p 的置信区间(置信度为 $1-\alpha$)

待估参数	其他参数	置信函数 G 的分布	近似的双侧置信区间
p	n 已知	$\dfrac{\left(\dfrac{\mu_n}{n}-p\right)}{\sqrt{\dfrac{p(1-p)}{n}}} \sim N(0,1)$	$\left(\dfrac{\mu_n}{n}-z_{\alpha/2}\sqrt{\dfrac{\dfrac{\mu_n}{n}\left(1-\dfrac{\mu_n}{n}\right)}{n}}, \dfrac{\mu_n}{n}+z_{\alpha/2}\sqrt{\dfrac{\dfrac{\mu_n}{n}\left(1-\dfrac{\mu_n}{n}\right)}{n}}\right)$

第 5 章　假设检验

5.1　问题提法和基本概念

5.1.1　例子与问题提法

原假设 (零假设, 解消假设) 是 $\mathbb{R}^n$ 的一个子集, 其中 n 是未知参数的数量.

对立假设 (备择假设) 是原假设补集的子集.

检验统计量, 接受域, 否定域 (临界域), 临界值.

简单假设, 复合假设. 赘余参数 (多余参数, 讨厌参数).

双边 (右边, 左边) 备择假设, 双边 (单边) 检验.

5.1.2　功效函数

设总体分布中包含未知参数 $\theta_1, \theta_2, \cdots, \theta_k$, H_0 为原假设, H_1 为对立假设, Φ 是基于样本 $X_1, X_2, \cdots, X_n$ 而对 H_0 做的一个检验, 则其 **功效函数** 是:

$$\beta_\Phi(\theta_1, \theta_2, \cdots, \theta_n) = P_{\theta_1, \theta_2, \cdots, \theta_n}(\text{在检验 } \Phi \text{ 之下}, \ H_0 \text{ 被否定})$$

- 当 $(\theta_1, \theta_2, \cdots, \theta_n) \in H_0$ 时, 上式越小越好.
- 当 $(\theta_1, \theta_2, \cdots, \theta_n) \in H_1$ 时, 上式越大越好, 此时称为功效函数.

5.1.3　两类错误与假设检验思路

两类错误

1. 拒真: H_0 正确, 但被否定.
2. 取伪: H_0 错误, 但被接受.

$$\alpha_{1\Phi}(\theta_1, \theta_2, \cdots, \theta_k) = \begin{cases} \beta_\Phi(\theta_1, \theta_2, \cdots, \theta_k), & (\theta_1, \theta_2, \cdots, \theta_k) \in H_0, \\ 0 & (\theta_1, \theta_2, \cdots, \theta_k) \in H_1. \end{cases}$$

$$\alpha_{2\Phi}(\theta_1, \theta_2, \cdots, \theta_k) = \begin{cases} 0, & (\theta_1, \theta_2, \cdots, \theta_k) \in H_0, \\ 1 - \beta_\Phi(\theta_1, \theta_2, \cdots, \theta_k), & (\theta_1, \theta_2, \cdots, \theta_k) \in H_1. \end{cases}$$

奈曼-皮尔逊理论 的思路: 先保证第一类错误的概率不超过某个定值 α, 再使第二类错误的概率尽可能小.

5.1.4　检验水平与一致最优检验

H_0 的一个水平为 α 的检验 Φ:

$$\forall (\theta_1, \theta_2, \cdots, \theta_k) \in H_0 : \beta_\Phi(\theta_1, \theta_2, \cdots, \theta_k) \leq \alpha.$$

并使 α 尽可能小. 即固定第一类错误概率的原则.

- 若 Φ 是 H_0 的一个水平为 α_0 的检验, 则它也是水平为 α $(\forall \alpha > \alpha_0)$ 的检验.

假设检验问题 $H_0 : H_1$ 的一个水平为 α 的一致最优检验 Φ: 即对任何一个其它水平 α 的检验 g 有

$$\beta_\Phi(\theta_1,\theta_2,\cdots,\theta_k) \geq \beta_g(\theta_1,\theta_2,\cdots,\theta_k) \quad (\text{对任何 } (\theta_1,\theta_2,\cdots,\theta_k) \in H_1).$$

- 在总体分布只依赖于一个参数 θ, 而原假设 H_0 是 $\theta \leq \theta_0$ 或 $\theta \geq \theta_0$ 时, 一致最优检验存在.

5.2 重要参数检验

5.2.1 正态总体均值的检验

设 $X_1, X_2, \cdots, X_n$ 是从正态总体 $N(\mu, \sigma^2)$ 中抽出的样本.

原假设和对立假设分别为:

1. $H_0 : \mu \geq \mu_0,\ H_1 : \mu < \mu_0.$
2. $H_0' : \mu \leq \mu_0,\ H_1' : \mu > \mu_0.$
3. $H_0'' : \mu = \mu_0,\ H_1'' : \mu \neq \mu_0.$

1 方差已知

利用 $\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1).$

- $H_0 : \mu \geq \mu_0,\ H_1 : \mu < \mu_0.$
  - Φ : 当 $\overline{X} \geq \mu_0 - \dfrac{\sigma u_\alpha}{\sqrt{n}}$ 时接受原假设 H_0, 否则否定 H_0.
  - 功效函数 $\beta_\Phi(\mu) = \Phi\left(\dfrac{\sqrt{n}}{\sigma}(\mu_0 - \mu) - u_\alpha\right) \geq 1 - \beta.$
  - 欲使第二类错误概率足够小: $n \geq \dfrac{\sigma^2(u_\alpha + u_\beta)^2}{(\mu_0 - \mu_1)^2}.$
- $H_0' : \mu \leq \mu_0,\ H_1' : \mu > \mu_0.$
  - Φ : 当 $\overline{X} \leq \mu_0 + \dfrac{\sigma u_\alpha}{\sqrt{n}}$ 时接受原假设 H_0', 否则否定 H_0'.
  - 功效函数 $\beta_\Phi'(\mu) = 1 - \Phi\left(\dfrac{\sqrt{n}}{\sigma}(\mu_0 - \mu) + u_\alpha\right) \geq 1 - \beta.$
  - 欲使第二类错误概率足够小: $n \geq \dfrac{\sigma^2(u_\alpha + u_\beta)^2}{(\mu_0 - \mu_1)^2}.$
- $H_0'' : \mu = \mu_0,\ H_1'' : \mu \neq \mu_0.$
  - Φ'' : 当 $\left|\overline{X} - \mu_0\right| \leq \dfrac{\sigma u_{\alpha/2}}{\sqrt{n}}$ 时接受 H_0'', 否则否定 H_0''.
  - 功效函数 $\beta_\Phi''(\mu) = 2\Phi\left(\dfrac{\sqrt{n}}{\sigma}(\mu_0 - \mu) - u_{\alpha/2}\right) \geq 1 - \beta.$
  - 欲使第二类错误概率足够小: $n \geq \dfrac{\sigma^2(u_{\frac{\alpha}{2}} + u_{\frac{1+\beta}{2}})^2}{(\mu_0 - \mu_1)^2}.$

注

- 若 $\mu_0 - \dfrac{\sigma u_\alpha}{\sqrt{n}} \le \overline{X} \le \mu_0 + \dfrac{\sigma u_\alpha}{\sqrt{n}}$, 则既接受 H_0, 也接受 H_0'.
- $\varPhi$ 和 $\varPhi'$ 都是一致最优检验.
- $\varPhi''$ 不是一致最优检验, 并且 H_0'' 不存在一致最优检验.

2 方差未知

利用 $\dfrac{\overline{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$, 得到如下 t 检验:

- $H_0: \mu \ge \mu_0,\ H_1: \mu < \mu_0.$
  - $\varPsi$: 当 $\overline{X} \ge \mu_0 - \dfrac{S}{\sqrt{n}} t_{n-1}(\alpha)$ 时接受原假设 H_0, 否则否定 H_0.
  - 功效函数 $\beta_\varPsi(\mu,\sigma) = P_{\mu,\sigma}\left(\dfrac{\sqrt{n}}{S}(\overline{X}-\mu_0) < -t_{n-1}(\alpha)\right)$ 只依赖于 $\delta = \dfrac{\mu-\mu_0}{\sigma}$.
- $H_0': \mu \le \mu_0,\ H_1': \mu > \mu_0.$
  - $\varPsi'$: 当 $\overline{X} \le \mu_0 + \dfrac{S}{\sqrt{n}} t_{n-1}(\alpha)$ 时接受原假设 H_0', 否则否定 H_0'.
- $H_0'': \mu = \mu_0,\ H_1'': \mu \ne \mu_0.$
  - $\varPsi''$: 当 $\left|\overline{X}-\mu_0\right| \le \dfrac{S}{\sqrt{n}} t_{n-1}\left(\dfrac{\alpha}{2}\right)$ 时接受原假设 H_0'', 否则否定 H_0''.

注: 除非检验水平 $\alpha \ge \dfrac{1}{2}$, 否则 $\varPsi$ 和 $\varPsi'$ 都不是一直最优检验.

5.2.2 两个正态总体均值差的检验

设 $X_1, X_2, \cdots, X_{n_1}$ 和 $Y_1, Y_2, \cdots, Y_{n_2}$ 分别是从正态总体 $N(\mu_1, \sigma_1^2)$ 和 $N(\mu_2, \sigma_2^2)$ 中抽取的相互独立的样本.

原假设和对立假设分别为

1. $H_0: \mu_1 - \mu_2 \ge \mu_0,\ H_1: \mu_1 - \mu_2 < \mu_0.$
2. $H_0': \mu_1 - \mu_2 \le \mu_0,\ H_1': \mu_1 - \mu_2 > \mu_0.$
3. $H_0'': \mu_1 - \mu_2 = \mu_0,\ H_1'': \mu_1 - \mu_2 \ne \mu_0.$

1 方差已知

利用 $\overline{X}-\overline{Y} \sim N\left(\mu_1-\mu_2, \dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}\right)$, 记 $\mu_0 = \mu_1 - \mu_2$, $U = \dfrac{(\overline{X}-\overline{Y})-\mu_0}{\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}}$, 则

- g : 当 $U \ge -u_\alpha$ 时接受 H_0, 否则否定 H_0.
- g' : 当 $U \le u_\alpha$ 时接受 H_0', 否则否定 H_0'.
- g'' : 当 $|U| \le u_{\alpha/2}$ 时接受 H_0'', 否则否定 H_0''.

2 方差未知

对于 $\sigma_1^2=\sigma_2^2=\sigma^2$ 的情况, 利用 $T=\dfrac{(\overline{X}-\overline{Y})-\mu_0}{S_\omega\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}\sim t_{n_1+n_2-2},\ S_\omega^2=\dfrac{\mathrm{SS}_1+\mathrm{SS}_2}{n_1+n_2-2}$, 可得到如下

两样本 t 检验:

- h : 当 $T\geq -t_{n_1+n_2-2}(\alpha)$ 时接受 H_0, 否则否定 H_0.
- h' : 当 $T\leq t_{n_1+n_2-2}(\alpha)$ 时接受 H_0', 否则否定 H_0'.
- h'' : 当 $|T|\leq t_{n_1+n_2-2}\left(\dfrac{\alpha}{2}\right)$ 时接受 H_0'', 否则否定 H_0''.

显著性检验 (希望原假设被否定的检验)

5.2.3 正态分布方差的检验

设 $X_1,X_2,\cdots,X_n$ 是从正态总体 $N(\mu,\sigma^2)$ 中抽出的样本.

原假设和对立假设分别为:

1. $H_0:\sigma^2\geq\sigma_0^2,\ H_1:\sigma^2<\sigma_0^2$.
2. $H_0':\sigma^2\leq\sigma_0^2,\ H_1:\sigma^2>\sigma_0^2$.
3. $H_0'':\sigma^2=\sigma_0^2,\ H_1'':\sigma^2\neq\sigma_0^2$.

注: 方差检验的误差较大.

1 均值已知

利用 $\displaystyle\sum_{i=1}^n\frac{(X_i-\mu)^2}{\sigma^2}\sim\chi_n^2$,

- ϕ : 当 $\displaystyle\sum_{i=1}^n(X_i-\mu)^2\geq\sigma_0^2\chi_n^2(1-\alpha)$ 时接受 H_0, 否则否定 H_0.
- ϕ' : 当 $\displaystyle\sum_{i=1}^n(X_i-\mu)^2\leq\sigma_0^2\chi_n^2(\alpha)$ 时接受 H_0', 否则否定 H_0'.
- ϕ'' : 当 $\displaystyle\sigma_0^2\chi_n^2\left(1-\frac{\alpha}{2}\right)\leq\sum_{i=1}^n(X_i-\mu)^2\leq\sigma_0^2\chi_n^2\left(\frac{\alpha}{2}\right)$ 时接受 H_0'', 否则否定 H_0''.

2 均值未知

利用 $\mathrm{SS}\sim\sigma^2\chi_{n-1}^2$,

- φ : 当 $\mathrm{SS}\geq\sigma_0^2\chi_{n-1}^2(1-\alpha)$ 时接受 H_0, 否则否定 H_0.
- φ' : 当 $\mathrm{SS}\leq\sigma_0^2\chi_{n-1}^2(\alpha)$ 时接受 H_0', 否则否定 H_0'.
- φ'' : 当 $\sigma_0^2\chi_{n-1}^2\left(1-\dfrac{\alpha}{2}\right)\leq\mathrm{SS}\leq\sigma_0^2\chi_{n-1}^2\left(\dfrac{\alpha}{2}\right)$ 时接受 H_0'', 否则否定 H_0''.

5.2.π 两个正态分布方差商的检验

设 $X_1, X_2, \cdots, X_{n_1}$ 和 $Y_1, Y_2, \cdots, Y_{n_2}$ 分别是从正态总体 $N(\mu_1, \sigma^2)$ 和 $N(\mu_2, \sigma^2)$ 中抽取的相互独立的样本.

原假设和对立假设分别为

1. $H_0 : \sigma_1^2/\sigma_2^2 \geq \mu_0,\ H_1 : \sigma_1^2/\sigma_2^2 < \mu_0$.
2. $H_0' : \sigma_1^2/\sigma_2^2 \leq \mu_0,\ H_1' : \sigma_1^2/\sigma_2^2 > \mu_0$.
3. $H_0'' : \sigma_1^2/\sigma_2^2 = \mu_0,\ H_1'' : \sigma_1^2/\sigma_2^2 \neq \mu_0$.

1 均值已知

记 $U = \dfrac{\sum\limits_{i=1}^{n_1}(X_i-\mu_1)^2}{\sum\limits_{i=1}^{n_2}(Y_i-\mu_2)^2}$, 利用 $\dfrac{n_1\sigma_1^2}{n_2\sigma_2^2}\dfrac{1}{U} \sim F_{n_2,n_1}$, 下略.

2 均值未知

利用 $\dfrac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \sim F_{n_1-1,n_2-1}$,

- g : 当 $\dfrac{S_1^2}{S_2^2} \geq \dfrac{\sigma_1^2}{\sigma_2^2}F_{n_1-1,n_2-1}(1-\alpha) = \dfrac{\sigma_1^2}{\sigma_2^2}F^{-1}_{n_2-1,n_1-1}(\alpha)$ 时接受 H_0, 否则否定 H_0.
- g'' : 当 $\dfrac{\sigma_1^2}{\sigma_2^2}F^{-1}_{n_2-1,n_1-1}\left(\dfrac{\alpha}{2}\right) \leq \dfrac{S_1^2}{S_2^2} \leq \dfrac{\sigma_1^2}{\sigma_2^2}F_{n_1-1,n_2-1}\left(\dfrac{\alpha}{2}\right)$ 时接受 H_0'', 否则否定 H_0''.

5.2.4 指数分布参数的检验

1 普通检验

利用 $2n\lambda\overline{X} \sim \chi^2_{2n}$.

2 截尾寿命检验

2.1 定数截尾法

取 n 个元件做试验, 取 $r \in \mathbb{N}, r < n$, 进行到 r 个元件失效时为止, 所有元件工作的时间为

$T = Y_1 + \cdots + Y_r + (n-r)Y_r \sim \chi^2_r$.

2.2 定时截尾法

到时刻 T_0 为止, 所有元件工作的总时间为满足 $2\lambda T \dot{\sim} \chi^2_{2u+1}$, 其中 u 是已经失效的元件个数.

5.2.5 二项分布参数的检验

普通检验

$H_0: p \le p_0,\ H_1: p > p_0.$

φ : 当 $X < C$ 时接受 H_0, 否则否定 H_0.

$$\beta_\varphi(p) = P_p(X > C) = 1 - \sum_{i=0}^{C} \binom{n}{i} p^i (1-p)^{n-i}.$$

一般可查表, 对于较大的样本, 难以查表, 可以使用大样本法.

随机化检验

检验 φ 的 **操作特征函数** (**OC 函数**)

符号检验

非参数检验

5.2.6 泊松分布参数的检验

利用如下性质:

1. 泊松分布和的函数 (可加性)
   $X_1 \sim P(\lambda_1),\ X_2 \sim P(\lambda_2) \quad \Rightarrow \quad X_1 + X_2 \sim P(\lambda_1 + \lambda_2).$
2. $P_\lambda(X \le k) = \sum\limits_{i=0}^{k} \dfrac{\mathrm{e}^{-\lambda}\lambda^i}{i!} = \displaystyle\int_\lambda^{+\infty} \frac{\mathrm{e}^{-t}t^k}{k!}\,\mathrm{d}t = K_{2k+2}(2\lambda)$. (卡方分布函数)
3. 若有一批零件寿命服从指数分布, 固定一个时间 $T > 0$, 让一个元件从时刻 0 开始工作, 每当这个元件坏了的时候马上用一个新的替换, 则到 T 时替换的次数 $X \sim P(\lambda T)$, 即 $P(X = n) = \dfrac{\mathrm{e}^{-\lambda T}(\lambda T)^n}{n!}$.
   - 若有 n 个元件同时开始工作, 每个元件损坏即替换, 则 $X \sim P(n\lambda T)$.

5.2.7 大样本检验

贝伦斯—费歇尔问题

$$T = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{S_1^2/n_1 + S_2^2/n_2}} \dot\sim \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \sim N(0,1).$$

二项分布参数检验

$$\frac{X - np}{\sqrt{np(1-p)}} \dot\sim N(0,1).$$

一般分布参数检验

$$\frac{\sum X - nE(X)}{\sqrt{n\operatorname{Var}(X)}} = \frac{\sum X - n\mu}{\sqrt{n}\sigma} \sim N(0,1)$$

5.2.8　贝叶斯方法

基本思路

若原假设的条件概率大于对立假设的条件概率:

$$P(H_0 \mid X_1, X_2, \cdots, X_n) > P(H_1 \mid X_1, X_2, \cdots, X_n)$$

则接受原假设.

正态分布的区间检验

设 $X_1, X_2, \cdots, X_n$ 为抽自正态总体 $N(\theta, \sigma^2)$ 的样本, 其中 θ 和 σ^2 都未知, 考虑检验问题

$$H_0 : a \leq \theta \leq b, \quad H_1 : \theta < a \vee \theta > b \ (a < b).$$

给 (θ, σ) 以广义先验密度 σ^{-1}, 则在得知样本 $X_1, X_2, \cdots, X_n$ 后 (θ, σ) 的后验密度为

$$h(\theta, \sigma) = C_n \sigma^{-(n+1)} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \theta)^2\right], \quad (\theta \in \mathbb{R},\ \sigma > 0)$$

$$C_n = \left(\int_0^{+\infty} \mathrm{d}\sigma \int_{-\infty}^{+\infty} \sigma^{-(n+1)} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \theta)^2\right] \mathrm{d}\sigma\right)^{-1}$$

从而得到 θ 的边缘后验密度为

$$\begin{aligned} h_\theta(\theta) &= C_n \int_0^{+\infty} \sigma^{-(n+1)} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \theta)^2\right] \mathrm{d}\sigma \\ &= D_n \left[\sum_{i=1}^{n}(X_i - \theta)^2\right]^{-\frac{n}{2}} = D_n\left[\mathrm{SS} + n(\overline{X} - \theta)^2\right]^{-\frac{n}{2}} \\ &= D_n \mathrm{SS}^{-n}\left[1 + \frac{n(\overline{X} - \theta)^2}{\mathrm{SS}}\right]^{-\frac{n}{2}} = E_n\left[1 + \frac{n(\overline{X} - \theta)^2}{(n-1)S^2}\right]^{-\frac{n}{2}} \end{aligned}$$

令 $\theta^* = \dfrac{\sqrt{n}(\theta - \overline{X})}{S}$, 则比较 t 分布密度函数知 $\theta^* \sim t_{n-1}$, 即

$$f_{\theta^*} = F_n\left(1 + \frac{\theta^{*2}}{n-1}\right)^{-\frac{n}{2}} = \frac{\left(1 + \dfrac{\theta^{*2}}{n-1}\right)^{-\frac{n}{2}}}{\mathrm{B}\left(\dfrac{n-1}{2}, \dfrac{1}{2}\right)\sqrt{n-1}}$$

由此可以得到 ⭐

$$
\begin{aligned}
P(H_0 \mid X_1, \cdots, X_n) &= P(a \leq \theta \leq b \mid X_1, \cdots, X_n) \\
&= P\left(\frac{\sqrt{n}(a-\overline{X})}{S} \leq \theta^* \leq \frac{\sqrt{n}(b-\overline{X})}{S} \mid X_1, \cdots X_n\right) \\
&= T_{n-1}\left(\frac{\sqrt{n}(b-\overline{X})}{S}\right) - T_{n-1}\left(\frac{\sqrt{n}(a-\overline{X})}{S}\right)
\end{aligned}
$$

信仰推断法: 信仰分布, 信仰概率.

5.3 拟合优度检验

5.3.1 理论分布完全已知且只取有限个值的情况

问题提法

1. H_0：总体 X 的分布律为 $F_0(x)$.
2. H_0：总体 X 的分布律为 $P(X=x_i)=p_i,\ i=1,2,\cdots$.
3. H_0：总体 X 的概率密度函数为 $f_0(x)$.

对于原假设 $H_0: P(X=a_i)=p_i\ (i=1,2,\cdots,k)$, 设对 X 进行了足够多的 n 次实验, $X_1, X_2, \cdots, X_n$ 中等于 a_i 的个数记作 ν_i, 称为 a_i 这个 "类" 的 **经验值** 或 **观察值**, 其近似于**理论值** np_i.

皮尔逊统计量 $\chi^2 = \sum\limits_{i=1}^{k} \dfrac{(np_i-\nu_i)^2}{np_i} = \sum\limits_{i=1}^{k} \dfrac{\nu_i^2}{np_i} - n$.

定理 3.1 ⭐ 若原假设 H_0 成立, 则 $\chi^2 \dot{\sim} \chi^2_{k-1}$.

拟合优度 $p(Z_0)=P(Z \geq Z_0 \mid H_0) \approx 1-K_{k-1}(Z_0)$.

- 统计上的显著性不等于实用上的重要性.

5.3.2 理论分布只含有限个值但不完全已知的情况

设总体 $X \sim P(X=a_i)=p_i(\theta_1, \cdots, \theta_r)\ (i=1,2,\cdots,k;\ r \leq k-2)$, 并对 X 进行了足够多的 n 次观察, 记 $X_1, X_2, \cdots, X_n$ 中等于 a_i 的个数为 ν_i, 原假设为 H_0：总体分布对 $\theta_1^0, \theta_2^0, \cdots, \theta_r^0$ 成立.

定理 3.2 ⭐ 在一定条件下, 若原假设 H_0 成立, 使用极大似然法估计出 $(\hat{\theta}_1, \hat{\theta}_2, \cdots, \hat{\theta}_r)$, 并由此算出 $p_i = p_i(\hat{\theta}_1, \hat{\theta}_2, \cdots, \hat{\theta}_r)$, 则 $\chi^2 \dot{\sim} \chi^2_{k-1-r}$.

- 此时拟合优度约为 $p(Z_0)=1-K_{k-1-r}(Z_0)$.
- 一般取 $n \geq 50$, 且每一个理论值 np_i 的值都不小于 5.

5.3.3 对列联表的应用

$B\backslash A$	1	2	$\cdots$	i	$\cdots$	a	和
1	n_{11}	n_{12}	$\cdots$	n_{i1}	$\cdots$	n_{a1}	$n_{\cdot 1}$
2	n_{12}	n_{22}	$\cdots$	n_{i2}	$\cdots$	n_{a2}	$n_{\cdot 2}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
j	n_{1j}	n_{2j}	$\cdots$	n_{ij}	$\cdots$	a_{aj}	$n_{\cdot j}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
b	n_{1b}	n_{2b}	$\cdots$	n_{ib}	$\cdots$	n_{ab}	$n_{\cdot b}$
和	$n_{1\cdot}$	$n_{2\cdot}$	$\cdots$	$n_{i\cdot}$	$\cdots$	$n_{a\cdot}$	n

记 $P(A=i)=u_i,\, P(B=j)=v_j,\, P(A=i,\, B=j)=p_{ij}$.

$H_0:A$ 和 B 独立, 即 $p_{ij}=u_iv_j$.

$$L=\prod_{i=1}^{a}\prod_{j=1}^{b}(u_iv_j)^{n_{ij}}=\prod_{i=1}^{a}u_i^{n_{i\cdot}}\prod_{j=1}^{b}u_j^{n_{\cdot j}}$$

$$\begin{cases} 0=\dfrac{\partial\ln L}{\partial u_i}=\dfrac{n_{i\cdot}}{u_i}-\dfrac{n_{a\cdot}}{u_a} \\ 0=\dfrac{\partial\ln L}{\partial v_j}=\dfrac{n_{\cdot j}}{v_j}-\dfrac{n_{a\cdot}}{v_b} \\ \displaystyle\sum_{i=1}^{a}n_{i\cdot}=\sum_{j=1}^{b}n_{\cdot j}=n \end{cases} \quad\Rightarrow\quad \begin{cases} \hat{u}_i=\dfrac{n_{i\cdot}}{n} \\ \hat{v}_j=\dfrac{n_{\cdot j}}{n} \\ \hat{p}_{ij}=\dfrac{n_{i\cdot}n_{\cdot j}}{n^2} \end{cases}$$

自由度为 $k-1-r=ab-1-(a+b-2)=(a-1)(b-1)$.

$$Z=\sum_{i=1}^{a}\sum_{j=1}^{b}\frac{nn_{ij}-n_{i\cdot}n_{\cdot j}}{nn_{i\cdot}n_{\cdot j}}\dot{\sim}\chi^2_{(a-1)(b-1)}$$

当 $a=b=2$ 时称为 **四格表**, 此时有

$$Z=\frac{n(n_{11}n_{22}-n_{12}n_{21})^2}{n_{1\cdot}n_{2\cdot}n_{\cdot 1}n_{\cdot 2}}\dot{\sim}\chi^2_1$$

列联表可用于 **独立性检验** 或 **齐一性检验** 等.

5.3.4 总体分布为一般分布的情况

一般分布: 离散型有限个取值, 离散型无限个取值, 连续型.

$H_0:$ 总体分布为 $F(x)$ (或 $F(x;\theta_1^0,\theta_2^0,\cdots,\theta_r^0)$).

取划分 $-\infty=a_0<a_1<a_2<\cdots<a_{k-1}<a_k=+\infty$, 则区间 $I_i=(a_{i-1},a_i]$ 的概率为 $p_i(\theta_1,\cdots,\theta_r)=F(a_i;\theta_1,\cdots,\theta_r)-F(a_{i-1};\theta_1,\cdots,\theta_r),\quad(i=1,2,\cdots,k)$.

记 H_0' : 对某一组值 $(\theta_1^0,\theta_2^0,\cdots,\theta_r^0)$, 总体在区间 I_i 内的概率为 $p_i(\theta_1^0,\theta_2^0,\cdots,\theta_r^0)$.

记总体 X 的样本值 x_i 落在 I_i 内的个数为 n_i, 称为 **实际频数**, 其频率为 $f_i=\dfrac{n_i}{n}$, **理论频数** 为 np_i, 则检验方式同有限取值的无参数或有参数情况相同.

- 由于积分等使得表达式的计算较为困难, 实际中可不采取极大似然估计, 而使用矩估计近似替代.
- 如果初始数据就只给出了各区间的数量, 而无精确的数据, 可使用区间的中点近似计算矩估计, 如:

$$\hat{\mu} = \frac{1}{n}\sum_i m_i\nu_i, \quad \hat{\sigma^2} = \frac{1}{n}\sum_i \nu_i(m_i - \hat{\mu})^2.$$

表 8.3 正态总体均值、方差的假设检验(显著性水平为 α)

	原假设和备择假设	检验统计量	原假设 H_0 的拒绝域
单个正态总体	$H_0:\mu=\mu_0, H_1:\mu\neq\mu_0$ $H_0:\mu\leqslant\mu_0, H_1:\mu>\mu_0$ $H_0:\mu\geqslant\mu_0, H_1:\mu<\mu_0$ (σ^2 已知)	$Z=\dfrac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}\sim N(0,1)$	$\lvert Z\rvert\geqslant z_{\alpha/2}$ $Z\geqslant z_\alpha$ $Z\leqslant -z_\alpha$
	$H_0:\mu=\mu_0, H_1:\mu\neq\mu_0$ $H_0:\mu\leqslant\mu_0, H_1:\mu>\mu_0$ $H_0:\mu\geqslant\mu_0, H_1:\mu<\mu_0$ (σ^2 未知)	$T=\dfrac{\overline{X}-\mu_0}{S/\sqrt{n}}\sim t(n-1)$	$\lvert T\rvert\geqslant t_{\alpha/2}(n-1)$ $T\geqslant t_\alpha(n-1)$ $T\leqslant -t_\alpha(n-1)$
	$H_0:\sigma^2=\sigma_0^2, H_1:\sigma^2\neq\sigma_0^2$ $H_0:\sigma^2\leqslant\sigma_0^2, H_1:\sigma^2>\sigma_0^2$ $H_0:\sigma^2\geqslant\sigma_0^2, H_1:\sigma^2<\sigma_0^2$ (μ 未知)	$\chi^2=\dfrac{(n-1)S^2}{\sigma_0^2}\sim\chi^2(n-1)$	$\chi^2\leqslant\chi^2_{1-\alpha/2}(n-1)$ 或 $\chi^2\geqslant\chi^2_{\alpha/2}(n-1)$ $\chi^2\geqslant\chi^2_\alpha(n-1)$ $\chi^2\leqslant\chi^2_{1-\alpha}(n-1)$
两个正态总体	$H_0:\mu_1-\mu_2=\delta, H_1:\mu_1-\mu_2\neq\delta$ $H_0:\mu_1-\mu_2\leqslant\delta, H_1:\mu_1-\mu_2>\delta$ $H_0:\mu_1-\mu_2\geqslant\delta, H_1:\mu_1-\mu_2<\delta$ (σ_1^2,σ_2^2 已知)	$Z=\dfrac{\overline{X}-\overline{Y}-\delta}{\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}}\sim N(0,1)$	$\lvert Z\rvert\geqslant z_{\alpha/2}$ $Z\geqslant z_\alpha$ $Z\leqslant -z_\alpha$
	$H_0:\mu_1-\mu_2=\delta, H_1:\mu_1-\mu_2\neq\delta$ $H_0:\mu_1-\mu_2\leqslant\delta, H_1:\mu_1-\mu_2>\delta$ $H_0:\mu_1-\mu_2\geqslant\delta, H_1:\mu_1-\mu_2<\delta$ ($\sigma_1^2=\sigma_2^2=\sigma^2$ 未知)	$\dfrac{\overline{X}-\overline{Y}-\delta}{S_\omega\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}\sim t(n_1+n_2-2)$	$\lvert T\rvert\geqslant t_{\alpha/2}(n-1)$ $T\geqslant t_\alpha(n_1+n_2-2)$ $T\leqslant -t_\alpha(n_1+n_2-2)$
	$H_0:\sigma_1^2=\sigma_2^2, H_1:\sigma_1^2\neq\sigma_2^2$ $H_0:\sigma_1^2\leqslant\sigma_2^2, H_1:\sigma_1^2>\sigma_2^2$ $H_0:\sigma_1^2\geqslant\sigma_2^2, H_1:\sigma_1^2<\sigma_2^2$ (μ_1,μ_2 均未知)	$S_1^2/S_2^2\sim F(n_1-1,n_2-1)$	$F\leqslant F_{1-\alpha/2}(n_1-1,n_2-1)$ 或 $F\geqslant F_{\alpha/2}(n_1-1,n_2-1)$ $F\geqslant F_\alpha(n_1-1,n_2-1)$ $F\leqslant F_{1-\alpha}(n_1-1,n_2-1)$

常用标准正态分布值:

$$\begin{aligned} &z_{0.1}=1.281552 && z_{0.05}=1.644854 && z_{0.025}=1.959964\\ &z_{0.01}=2.326348 && z_{0.005}=2.575829 && z_{0.0025}=2.807034\\ &z_{0.001}=3.090232 \end{aligned}$$

更详细的统计分布表可以参考概统笔记附录.

第 6 章　回归、相关与方差分析

6.1　基本概念

自变量 (预报因子), 因变量 (预报量).

(理论 / 经验) 回归函数 / 回归方程.

非参数回归, 参数回归.

线性回归, 非线性回归.

> 回归分析一词的由来的本质原因, 可以参考我之前的一次 PPT 汇报.

回归分析的步骤:

1. 建立回归模型.
2. 估计参数, 确定回归方程.
3. 检验与评价回归方程.
4. 利用回归方程进行预测和控制.

6.2　一元线性回归

- 一元线性回归模型: $Y=b_0+b_1X+\varepsilon$.
  - 对于每一个 X_i, 有 $Y_i=b_0+b_1X_i+\varepsilon_i$.
  - 一般认为 ε_i 独立同分布 $N(0,\sigma^2)$.
- 对模型进行中心化: $Y=\beta_0+\beta_1(X-\overline{X})+\varepsilon$.
  - 其中 $\beta_1=b_1,\ \beta_0=b_0+b_1\overline{X}$.
  - 预测值记为 $\hat{Y}_i=\hat{\beta}_0+\hat{\beta}_1(X_i-\overline{X})$.

注

- 在回归分析时, 将 X_i 看作非随机常数. 为了强调这点, 一些资料上使用小写字母表示.
- 中心化是为了方便讨论. 许多资料上将未中心化的系数记为 β_i, 请注意区分公式中的异同;

6.2.1　最小二乘法

1 回归参数的点估计

待定系数的预测值记为 $\hat{Y}_{\mathrm{i}}=\alpha_0+\alpha_1(X_i-\overline{X})$, 并用 $Q(\alpha_0,\alpha_1)=\sum\limits_{i=1}^{n}(Y_i-\hat{Y})^2$ 衡量误差, 将最值点记为 $(\hat{\beta}_0,\hat{\beta}_1)$, 则有:

$$\begin{cases} \dfrac{\partial Q}{\partial \alpha_0} = 0, \\ \dfrac{\partial Q}{\partial \alpha_1} = 0. \end{cases} \quad \Rightarrow \quad \begin{cases} \hat{\beta}_0 = \overline{Y}, \\ \hat{\beta}_1 = \dfrac{\displaystyle\sum_{i=1}^{n}(X_i - \overline{X})Y_i}{\displaystyle\sum_{i=1}^{n}(X_i - \overline{X})^2} \end{cases} \quad \Rightarrow \quad \begin{cases} \hat{b}_1 = \dfrac{\displaystyle\sum_{i=1}^{n}(X_i - \overline{X})Y_i}{\displaystyle\sum_{i=1}^{n}(X_i - \overline{X})^2} \\ \hat{b}_0 = \overline{Y} - \hat{b}_1\overline{X} \end{cases}$$

- 其中分子的 Y_i 也可写为 $(Y_i - \overline{Y})$, 不影响结果, 此时一般简记为 $\hat{\beta}_1 = \dfrac{S_{xy}}{S_{xx}}$.

  注意这里的 S_{xy} 等, 都未开根号.
- 由上, 我们有 $\hat{Y} - \overline{Y} = \hat{b}_1(X - \overline{X})$.
- 除了用 Q, 还可以利用 $Y_i = b_0 + b_1X_i + \varepsilon_i \sim N(\beta_0 + \beta_1X_i, \sigma^2)$ 求解最大似然估计, 结果是一样的.

2 参数估计量的性质

1. 无偏估计
   1. $E(\hat{\beta}_0) = \beta_0$.
   2. $E(\hat{\beta}_1) = \beta_1$.
2. 方差
   1. $\mathrm{Var}(\hat{\beta}_0) = \dfrac{\sigma^2}{n}$.
   2. $\mathrm{Var}(\hat{\beta}_1) = \sigma^2 \bigg/ \displaystyle\sum_{i=1}^{n}(X_i - \overline{X})^2 = \dfrac{\sigma^2}{S_{xx}}$.
3. 协方差
   1. $\hat{\beta}_0 - E(\hat{\beta}_0) = \overline{\varepsilon}$.
   2. $\hat{\beta}_1 - E(\hat{\beta}_1) = \dfrac{\displaystyle\sum_{i=1}^{n}(X_i - \overline{X})\varepsilon_i}{\displaystyle\sum_{i=1}^{n}(X_i - \overline{X})^2}$.
   3. $\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) = 0$.
4. 如果 ε_i 独立同分布 $N(0, \sigma^2)$, 则 Y_i 服从正态分布, 于是 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 服从正态分布, 因此可由不相关推出独立.
   1. $\hat{\beta}_0 \sim N\left(\beta_0, \dfrac{\sigma^2}{n}\right)$.
   2. $\hat{\beta}_1 \sim N\left(\beta_1, \dfrac{\sigma^2}{S_{xx}}\right)$.
   3. $\hat{b}_0 \sim N\left(b_0, \left(\dfrac{1}{n} + \dfrac{\overline{X}^2}{S_{xx}}\right)\sigma^2\right)$.
5. $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 是最小方差无偏估计.

3 误差与残差的性质

- 概念
  - 误差 $\varepsilon_i = Y_i - b_0 - b_1 X_i$.
  - 残差 $\delta_i = Y_i - \hat{b}_0 - \hat{b}_1 X_i = Y_i - \hat{Y}_i$.
  - 残差平方和 $Q_\delta = \sum\limits_{i=1}^n \delta_i^2$.

> 有些资料中将误差与残差都用 ε 表示, 有些资料虽然区分了, 但将误差记为 e, 或者将残差记为 e. 为减少歧义, 这里一律采用上述记号.

- 性质
  - 理论值
    - $E(\delta_i) = 0$.
    - $\mathrm{Var}(\delta_i) = \left[n - 1 - (X_i - \overline{X})^2 \middle/ \sum\limits_{j=1}^n (X_j - \overline{X})^2\right] \sigma^2$.
  - 观测值
    - $\sum\limits_{i=1}^n \delta_i = 0$.
    - $\sum\limits_{i=1}^n X_i \delta_i = 0$.

    > 由最小二乘法中的偏导为零即得; 误差无此性质.
  - 易于计算的形式

$$\begin{aligned} Q_\delta = \sum_{i=1}^n \delta_i^2 &= \sum_{i=1}^n (Y_i - \overline{Y})^2 - \hat{\beta}_1 \sum_{i=1}^n (X_i - \overline{X}) Y_i \\ &= \sum_{i=1}^n Y_i^2 - n\overline{Y}^2 - \hat{\beta}_1 \sum_{i=1}^n (X_i - \overline{X}) Y_i \\ &= S_{yy} - \hat{\beta}_1 S_{xy} = S_{yy} - \frac{S_{xy}^2}{S_{xx}} \end{aligned}$$

    > 欲求出回归参数、误差方差的无偏估计，并对回归参数进行区间估计或假设检验，只需求出 S_{yy}, S_{xy} 和 S_{xx} 即可.

- 定理
  - 误差方差的无偏估计: $\hat{\sigma}^2 = \dfrac{1}{n-2} \sum\limits_{i=1}^n \delta_i^2 = \dfrac{Q_\delta}{n-2}$.
  - $\dfrac{Q_\delta}{\sigma^2} \sim \chi_{n-2}^2$, 且 $\overline{Y}$, $\hat{\beta}_1$, Q_ε 相互独立.

6.2.2 区间估计与假设检验

1 平方和表示

$\mathrm{SS}_{总} = \mathrm{SS}_{回} + \mathrm{SS}_{误}$.

- 总离差平方和 $\mathrm{SS}_{总} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = S_{yy}$.
- 回归平方和 $\mathrm{SS}_{回} = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 = \hat{\beta}_1 S_{xy} = \dfrac{S_{xy}^2}{S_{xx}}$.
- 误差平方和 $\mathrm{SS}_{误} = 2\sum_{i=1}^{n}(Y_i - \hat{Y}_i)(\hat{Y}_i - \overline{Y}) + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = Q_\delta$.

2 常用统计量

假定 ε_i 独立同分布 $N(0, \sigma^2)$, 则下述参数均服从正态分布.

- β_1.
  - $\dfrac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{S_{xx}}} \sim N(0,1)$.
  - $\dfrac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}$.
  - **t 统计量**: $\dfrac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/\sqrt{S_{xx}}} \sim t_{n-2}$.
  - **F 统计量**: $\dfrac{(\hat{\beta}_1 - \beta_1)^2}{\hat{\sigma}^2/S_{xx}} \sim F_{1,n-2}$.
- β_0.
  - $\dfrac{\hat{\beta}_0 - \beta_0}{\sigma/\sqrt{n}} \sim N(0,1)$.
  - **t 统计量**: $\dfrac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}/\sqrt{n}} \sim t_{n-2}$.
  - **F 统计量**: $\dfrac{(\hat{\beta}_0 - \beta_0)^2}{\hat{\sigma}/n} \sim F_{1,n-2}$.
- $m(x) = \beta_0 + \beta_1(x - \overline{X})$.
  - $\hat{m}(x) = \hat{\beta}_0 + \hat{\beta}_1(x - \overline{X})$.
  - $E[\hat{m}(x)] = m(x)$.
  - $\mathrm{Var}[\hat{m}(x)] = \lambda(x) = \sigma^2\left[\dfrac{1}{n} + \dfrac{(x-\overline{X})^2}{S_{xx}}\right]$.
  - $\hat{\lambda}(x) = \hat{\sigma}^2\left[\dfrac{1}{n} + \dfrac{(x-\overline{X})^2}{S_{xx}}\right]$.
  - $\dfrac{\hat{m}(x) - m(x)}{\sqrt{\lambda(x)}} \sim N(0,1)$.

- $\dfrac{\hat{m}(x)-m(x)}{\sqrt{\hat{\lambda}(x)}} \sim t_{n-2}$.
  - 越靠近样本中心点处预测越精确.
- $Y_0 = m(x_0) + \varepsilon_0$.
  - $\eta = Y_0 - \hat{m}(x)$.
  - $E(\eta) = 0$.
  - $\operatorname{Var}(\eta) = \sigma^2 \left[1 + \dfrac{1}{n} + \dfrac{x-\overline{X}}{S_{xx}}\right]$.
  - $\hat{\sigma}_\eta^2 = \hat{\sigma}^2 \left[1 + \dfrac{1}{n} + \dfrac{x-\overline{X}}{S_{xx}}\right]$.
  - $\dfrac{\eta}{\hat{\sigma}_\eta} \sim t_{n-2}$.

3 显著性检验

区间估计和假设检验直接利用常用统计量中的结论即可.

比较特殊且实用的是 $H_0 : \beta_1 = 0,\ H_1 : \beta_1 \neq 0$, 这是对线性假设的检验, 又称为 **回归显著性检验**.
代入即得显著性水平 α 下 H_0 的拒绝域为:

- **t 统计量**: $|T| = \dfrac{\left|\hat{\beta}_1\right|\sqrt{S_{xx}}}{\hat{\sigma}} = \dfrac{\sqrt{n-2}\,|S_{xy}|}{\sqrt{S_{xx}S_{yy}-S_{xy}^2}} \geq t_{n-2}\left(\dfrac{\alpha}{2}\right)$.
- **F 统计量**: $F = T^2 = \dfrac{(n-2)\mathrm{SS}_{回}}{\mathrm{SS}_{误}} = \dfrac{\mathrm{MS}_{回}}{\mathrm{MS}_{误}} \geq F_{1,n-2}(\alpha)$.
- **R 统计量**: $|R| = \dfrac{|S_{xy}|}{\sqrt{S_{xx}S_{yy}}} = \sqrt{\dfrac{\mathrm{SS}_{回}}{\mathrm{SS}_{总}}} \geq r_{n-2}(\alpha)$.

一元线性回归的方差分析表

	自由度 (df)	平方和 (SS)	均方 (MS)	F
回归分析	1	$\mathrm{SS}_{回}$	$\mathrm{MS}_{回}$	
残差	$n-2$	$\mathrm{SS}_{误}$	$\mathrm{MS}_{误}$	$\dfrac{\mathrm{MS}_{回}}{\mathrm{MS}_{误}}$
总计	$n-1$	$\mathrm{SS}_{总}$		

相关系数绝对值	$\|R\|$
相关系数的平方	R^2

6.2.3 预测与控制

1 点预测

点预测值 $\hat{Y}_0=\hat{b}_0+\hat{b}_1x_0$ 与随机变量 $Y_0=b_0+b_1x_0+\varepsilon_0$ 有相同的数学期望.

2 区间预测

$\hat{Y}_0$ 和 δ_0 都可以用 Y_i 线性表示, 因此均为正态分布, 并有

$$\begin{aligned}\hat{Y}_0&=\overline{Y}+\hat{\beta}_1(X_0-\overline{X})\\&=\sum_{i=1}^n\left(\frac{1}{n}+\frac{(X_i-\overline{X})(X_0-\overline{X})}{S_{xx}}\right)Y_i\\&\sim N\left(b_0+b_1x_0,\left(\frac{1}{n}+\frac{(X_0-\overline{X})^2}{S_{xx}}\right)\sigma^2\right)\\\delta_0&=Y_0-\hat{Y}_0\\&\sim N\left(0,\left(1+\frac{1}{n}+\frac{(X_0-\overline{X})^2}{S_{xx}}\right)\sigma^2\right)\end{aligned}$$

结合 $\dfrac{(n-2)\hat{\sigma}^2}{\sigma^2}\sim\chi^2_{n-2}$, 因此有

$$\frac{\hat{Y}_0-Y_0}{\hat{\sigma}\sqrt{1+\dfrac{1}{n}+\dfrac{(X_0-\overline{X})^2}{S_{xx}}}}\sim t_{n-2}$$

3 控制

注意: 回归方程 $\hat{Y}=\hat{b}_0+\hat{b}_1X$ 不可推出 $\hat{X}=\dfrac{Y-\hat{b}_0}{\hat{b}_1}$, 但是可以估计 $X_0=\dfrac{\hat{Y}-\hat{b}_0}{\hat{b}_1}$, 即回归分析的控制作用.

6.3 多元线性回归

- **多元线性回归模型**: $Y=b_0+b_1X_1+b_2X_2+\cdots+b_pX_p+\varepsilon$.
  - 观测值: $Y_i=b_0+b_1X_{1i}+b_2X_{2i}+\cdots+b_pX_{pi}+\varepsilon$.
  - 一般认为 ε_i 独立同分布 $N(0,\sigma^2)$, 并且 X_i **非随机**.
- 对模型进行**中心化**: $Y=\beta_0+\beta_1X_1^*+\beta_2X_2^*+\cdots+\beta_pX_p^*+\varepsilon$.
  - 其中 $X_k^*=X_k-\overline{X}_k,\ \beta_k=b_k,\ k=1,\cdots,p,\ \beta_0=b_0+b_1\overline{X}_1+\cdots+b_p\overline{X}_p$.
  - 方便起见, 之后我会将未中心化时的 X_i 记为 x_i, 中心化后的 X_k^* 记为 X_k.
- 表示为矩阵的形式: $\boldsymbol{y}=\boldsymbol{X\beta}+\boldsymbol{\varepsilon}$, 其中 $\boldsymbol{X}$ 称为**设计矩阵**, 并且这里已中心化.

$$\boldsymbol{y}=\begin{pmatrix}Y_1\\Y_2\\\vdots\\Y_n\end{pmatrix},\quad \boldsymbol{\beta}=\begin{pmatrix}\beta_0\\\beta_1\\\vdots\\\beta_p\end{pmatrix},\quad \boldsymbol{X}=\begin{pmatrix}1 & X_{11} & X_{12} & \cdots & X_{1p}\\1 & X_{21} & X_{22} & \cdots & X_{2p}\\\vdots & \vdots & \vdots & & \vdots\\1 & X_{n1} & X_{n2} & \cdots & X_{np}\end{pmatrix},\quad \boldsymbol{\varepsilon}=\begin{pmatrix}\varepsilon_1\\\varepsilon_2\\\vdots\\\varepsilon_n\end{pmatrix}.$$

注: 设计矩阵的元素一般未中心化, 并且无上式中的第一列. 这样做是为了讨论方便.

6.3.1 最小二乘法

1 回归参数的点估计

记预测值为 $\hat{\boldsymbol{y}}=\boldsymbol{X\alpha}$, 用下式衡量误差

$$Q(\alpha_0,\alpha_1,\cdots,\alpha_p)=\sum_{i=1}^n(Y_i-\alpha_0-X_{1i}\alpha_1-\cdots-X_{pi}\alpha_p)^2$$

并将极值点记为 $\hat\beta_0,\hat\beta_1,\cdots,\hat\beta_p$, 令偏导为零则有

$$\begin{cases}\dfrac{\partial Q}{\partial\alpha_0}=0,\\\dfrac{\partial Q}{\partial\alpha_1}=0,\\\quad\cdots\cdots\\\dfrac{\partial Q}{\partial\alpha_p}=0.\end{cases}\Rightarrow\quad\begin{cases}n\hat\beta_0=\displaystyle\sum_{i=1}^n y_i,\\l_{11}\hat\beta_1+l_{12}\hat\beta_2+\cdots+l_{1p}\hat\beta_p=\displaystyle\sum_{i=1}^n X_{1i}y_i=S_{x_1y},\\\quad\cdots\cdots\\l_{p1}\hat\beta_1+l_{p2}\hat\beta_2+\cdots+l_{pp}\hat\beta_p=\displaystyle\sum_{i=1}^n X_{pi}y_i=S_{x_py}.\end{cases}$$

其中 $l_{uv}=\displaystyle\sum_{i=1}^n X_{ui}X_{vi}=S_{XuXv}$, $\boldsymbol{L}=\boldsymbol{X}^{\mathrm T}\boldsymbol{X}$.

于是方程组化为*正则方程*: $\boldsymbol{L}\hat{\boldsymbol{\beta}}=\boldsymbol{X}^{\mathrm T}\boldsymbol{y}_{(n)}$.

记 $\boldsymbol{C}=\boldsymbol{L}^{-1}$, 则有 $\hat{\boldsymbol{\beta}}=(\boldsymbol{X}^{\mathrm T}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm T}\boldsymbol{y}=\boldsymbol{C}\boldsymbol{X}^{\mathrm T}\boldsymbol{y}$.

类似的, 这也是最大似然估计量.

2 参数估计量的性质

- $\hat{\boldsymbol{\beta}}$ 是 $\boldsymbol{\beta}$ 的无偏估计.
- $\mathrm{Cov}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\beta}})=(\boldsymbol{X}^{\mathrm T}\boldsymbol{X})^{-1}\sigma^2=\boldsymbol{C}\sigma^2$.
  - 方差
    - $\mathrm{Var}(\hat\beta_0)=\dfrac{\sigma^2}{n}$.
    - $\mathrm{Var}(\hat\beta_i)=c_{ij}\sigma^2$.
  - 协方差
    - $\mathrm{Cov}(\hat\beta_0,\hat\beta_i)=0\ (i=1,2,\cdots,p)$.
    - $\mathrm{Cov}(\hat\beta_i,\hat\beta_j)=c_{ij}\sigma^2$.

3 误差与残差的性质

概念同一元线性回归.

- 性质
  - 残差的性质
    - $\sum_{i=1}^{n} \delta_i = 0.$
    - $\sum_{i=1}^{n} X_{ki}\delta_i = 0.$

    > 由最小二乘法中的偏导为零即得.
  - 残差平方和的易于计算的形式

$$Q_\delta = S_{yy} - (\hat{\beta}_1 S_{x_1y} + \hat{\beta}_2 S_{x_2y} + \cdots + \hat{\beta}_p S_{x_py}).$$

- 定理
  - $\hat{\sigma}^2 = \dfrac{Q_\delta}{n-p-1}$ 是 σ^2 的无偏估计.
  - $\dfrac{Q_\delta}{\sigma^2} \sim \chi^2_{n-p-1}.$
  - 注意: 这里参数的数量为 $p+1$.

6.3.2 区间估计与假设检验

显著性检验的假设为 $H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0.$

- 统计量
  - $\dfrac{\mathrm{SS}_{回}}{\sigma^2} \sim \chi^2_p.$
  - $\dfrac{\mathrm{SS}_{误}}{\sigma^2} \sim \chi^2_{n-p-1}.$
- 拒绝域
  - F 统计量: $F = \dfrac{\mathrm{SS}_{回}/p}{\mathrm{SS}_{误}/(n-p-1)} = \dfrac{\mathrm{MS}_{回}}{\mathrm{MS}_{误}} \geq F_{p,n-p-1}(\alpha).$
  - R 统计量: $|R| = \sqrt{\dfrac{\mathrm{SS}_{回}}{\mathrm{SS}_{总}}} \geq r_{n-p-1}(\alpha).$

多元线性回归的方差分析表

	自由度 (df)	平方和 (SS)	均方 (MS)	F
回归分析	p	$\mathrm{SS}_{回}$	$\mathrm{MS}_{回}$	
残差	$n-p-1$	$\mathrm{SS}_{误}$	$\mathrm{MS}_{误}$	$\dfrac{\mathrm{MS}_{回}}{\mathrm{MS}_{误}}$
总计	$n-1$	$\mathrm{SS}_{总}$		

相关系数绝对值	$\lvert R\rvert$
相关系数的平方	R^2

6.3.3 预测与控制

1. 点预测

   $E(\hat{Y})=E(Y)$, 故可用 $\hat{Y}=\hat{\beta}_0+\hat{\beta}_1X_1+\cdots+\hat{\beta}_pX_p$ 进行点预测.

2. 区间预测

$$T=\frac{\hat{Y}_0-Y_0}{\hat{\sigma}\sqrt{1+(1,\boldsymbol{x}_0^{\mathrm{T}})(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\begin{pmatrix}1\\x_0\end{pmatrix}}}\sim t_{n-p-1}.$$

6.4 相关分析

6.4.1 相关系数的估计和检验

6.4.2 偏相关

6.4.3 复相关

6.5 方差分析

试验指标.

可控因素, 不可控因素.

单因素方差分析, 多因素方差分析.

6.5.1 单因素方差分析

1 数学模型

因素 A 有 s 个 **水平** $A_1,A_2,\cdots,A_s$.

水平	观察值				样本均值
A_1	x_{11}	x_{12}	$\cdots$	x_{1n_1}	μ_1
A_2	x_{21}	x_{22}	$\cdots$	x_{2n_2}	μ_2
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
A_s	x_{s1}	x_{s2}	$\cdots$	x_{sn_s}	μ_s

- 记 $X_{ij}=\mu_i+\varepsilon_{ij}$, 假定 $\varepsilon_{ij}\sim N(0,\sigma^2)$ 且相互独立.

- 总次数 $n = \sum_{i=1}^{s} n_i$.
- 总平均 $\mu = \frac{1}{n}\sum_{i=1}^{s} n_i\mu_i$.

- 水平 A_i 的效应 $\delta_i = \mu_i - \mu$.
  - $X_{ij} = \mu + \delta_i + \varepsilon_{ij}$.
  - $\sum_{i=1}^{s} n_i\delta_i = 0$.
- 假设
  - $H_0 : \mu_1 = \mu_2 = \cdots = \mu_s$.
  - $H_0 : \delta_1 = \delta_2 = \cdots = \delta_s$.

2 总离差平方和

2.1 分解

- **总离差平方和** $\mathrm{SS_T} = \sum_{i=1}^{s}\sum_{j=1}^{n_i}(X_{ij} - \overline{X})^2$.
  - 组内均值 $\overline{X}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}$.
  - 总平均值 $\overline{X} = \frac{1}{n}\sum_{i=1}^{s}\sum_{j=1}^{n_i} X_{ij} = \frac{1}{n}\sum_{i=1}^{s} n_i\overline{X}_{i\cdot}$.
- 总离差平方和的 **分解** $\mathrm{SS_T} = \mathrm{SS_E} + \mathrm{SS_A}$.
  - **误差平方和** (组内方差) $\mathrm{SS_E} = \sum_{i=1}^{s}\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_{i\cdot})^2$.
  - **效应平方和** (组间方差) $\mathrm{SS_A} = \sum_{i=1}^{s}\sum_{j=1}^{n_i}(\overline{X}_{i\cdot} - \overline{X})^2$.

$$\mathrm{SS_A} = \sum_{i=1}^{s} n_i\overline{X}_{i\cdot}^2 - n\overline{X}^2 = \sum_{i=1}^{s}\frac{(\sum X_i)^2}{n_i} - \frac{(\sum X)^2}{n}.$$

2.2 统计特征

- $\mathrm{SS_E}$.
  - $\dfrac{\sum_{j=1}^{n_i}(X_ij - \overline{X}_{i\cdot})^2}{\sigma^2} \sim \chi^2_{n_i-1}$.
  - $\dfrac{\mathrm{SS_E}}{\sigma^2} \sim \chi^2_{n-s}$.
  - $E(\mathrm{SS_E}) = (n-s)\sigma^2$.
  - $\mathrm{Var}(\mathrm{SS_E}) = 2(n-s)\sigma^2$.
- $\mathrm{SS_A}$.

- $E(\mathrm{SS_A}) = (s-1)\sigma^2 + \sum_{i=1}^{s} n_i\delta_i^2$.
- 当 H_0 成立时, 有
  - $E(\mathrm{SS_A}) = (s-1)\sigma^2$.
  - $\frac{\mathrm{SS_A}}{\sigma^2} \sim \chi^2_{s-1}$.
  - $\mathrm{SS_A}$ 与 $\mathrm{SS_E}$ 相互独立.

3 拒绝域

拒绝 H_0, 即认为因素 A 各个水平的试验指标之间有显著差异.

$$F = \frac{\mathrm{SS_A}/(s-1)}{\mathrm{SS_E}/(n-s)} = \frac{\mathrm{MS_A}}{\mathrm{MS_E}} \geq F_{s-1,n-s}(\alpha).$$

SUMMARY 表

水平	观测数	总和	组均值	SS_i
水平 1	n_1	$\sum_{j=1}^{n_1} x_{1j}$	$\overline{x}_{1\cdot}$	SS_1
水平 2	n_1	$\sum_{j=1}^{n_2} x_{2j}$	$\overline{x}_{2\cdot}$	SS_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
水平 s	n_s	$\sum_{j=s}^{n_s} x_{sj}$	$\overline{x}_{s\cdot}$	SS_s

---

单因素方差分析表

差异源	平方和 (SS)	自由度 (df)	均方 MS	F	F 分位数 (F crit)
组间	$\mathrm{SS_A}$	$s-1$	$\mathrm{MS_A}$	$\frac{\mathrm{MS_A}}{\mathrm{MS_E}}$	$F_{s-1,n-s}(\alpha)$
组内	$\mathrm{SS_E}$	$n-s$	$\mathrm{MS_E}$		
总和	$\mathrm{SS_T}$	$n-1$			

上述参数的计算见总离差平方和.

4 参数估计

- $\hat{\mu}_i = \overline{X}_{i\cdot}$.
- $\hat{\mu} = \overline{X}$.
- $\hat{\delta}_i = \hat{\mu}_i - \hat{\mu}$.
- $\hat{\sigma}^2 = \mathrm{MS_E}$.