

# ch1 统计学前言

## 01 数据

	定类数据	定序数据	定量数据
分类	✓	✓	✓
排序		✓	✓
间距			✓
比值			✓

## 02 统计指标

### 平均数

#### 1. 算术平均数

性质：各变量值与算术平均数的差值1范数和二范数最小

$$\text{算术平均数} = \frac{\text{总体标志值总数}}{\text{总体单位数}}$$

#### 2. 加权算术平均数

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + \dots + x_n f_n}{f_1 + f_2 + \dots + f_n} = \frac{\sum x f}{\sum f}$$

#### 3. 调和平均数/倒数平均数：倒数算术平均数的倒数

$$\bar{x}_H = \frac{1}{\frac{\frac{1}{x_1} m_1 + \frac{1}{x_2} m_2 + \dots + \frac{1}{x_n} m_n}{m_1 + m_2 + \dots + m_n}} = \frac{m_1 + m_2 + \dots + m_n}{\frac{1}{x_1} m_1 + \frac{1}{x_2} m_2 + \dots + \frac{1}{x_n} m_n} = \frac{\sum m}{\sum \frac{1}{x} m}$$

#### 4. 几何平均数

应用：增长率

$$\bar{x}_G = \sqrt[n]{x_1 \cdot x_2 \cdot \dots \cdot x_n} = \sqrt[n]{\prod x^f}$$

#### 5. 加权几何平均数

$$\bar{x}_G = \sqrt[2f]{x_1^{f_1} \cdot x_2^{f_2} \cdot \dots \cdot x_n^{f_n}} = \sqrt[2f]{\prod_1^n x^f}$$

## 中位数

$M_e$ : 总体中各单位标志值按照大小顺序排列，处于中间位置的数

## 众数

$M_o$ :

$$Mo = L_m + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times d_m \quad (\text{下限公式})$$
$$Mo = U_m - \frac{\Delta_2}{\Delta_1 + \Delta_2} \times d_m \quad (\text{上限公式})$$

# 03 采样

## 01 简单随机抽样

### 01 要求

1. 要求总体个数有限
2. 从总体中逐个进行抽取
3. 不放回抽样
4. 总体中每一个个体被抽去的可能性相等

### 02 方法

1. 抽签法
2. 随机数法

### 03 样本条件

1. 独立性：相互独立
2. 代表性：每一个与总体有相同的分布

### 04 统计量

### 05 样本数字特征

1. 样本均值

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

2. 样本方差

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

3. 样本矩

$$K \text{ 阶原点矩: } \alpha_{nk} = \frac{1}{n} \sum_{i=1}^n X_i^k \quad (k = 1, 2, \dots) \quad \alpha_{n1} = \bar{X}$$

$$K \text{ 阶中心矩: } m_{nk} = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^k \quad (k = 1, 2, \dots) \quad m_{n2} = \frac{n-1}{n} s^2$$

一阶原点矩是样本均值

## 02 分层抽样

差异明显

每层样本数量与每层

## 03 整体抽样

整群抽样：将总体分成若干群，以群为抽样单位，对抽中的群所有基本单位调查

应用：质量检测

多阶段抽样：对抽中的群继续抽样

## 04 非随机的等距抽样

## 05 系统抽样

# 04 概率的基本概念

---

### 1. 频率与概率

频率

$$f_n(A) \triangleq n_A/n$$

### 2. Laplace 概率

### 3. 概率的公理化定义

非负性

规范性

可列可加性：对两两不相容事件

# 05 大数定律与中心极限定理

---

## 01 切比雪夫不等式

随机变量X的数学期望  $EX = \mu$ , 方差  $DX = \sigma^2$ , 则对任意

$$P\{|X - \mu| \geq \varepsilon\} \leq \frac{\sigma^2}{\varepsilon^2} \text{ 或 } P\{|X - \mu| < \varepsilon\} \geq 1 - \frac{\sigma^2}{\varepsilon^2}$$

## 02 相关性分析

相关系数

$$\begin{aligned} r &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \\ &= \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sqrt{(\sum_{i=1}^n x_i^2 - n(\bar{x})^2)(\sum_{i=1}^n y_i^2 - n(\bar{y})^2)}} \end{aligned}$$

$r > 0$  正相关

$r < 0$  负相关

$r$  的绝对值越接近 1, 表明两个变量的线形相关性越强

## 03 回归分析

## 04 区间估计

# ch3 概率抽样分布

## 条件概率

### 1. 定义

$$P(A | B) = \frac{P(AB)}{P(B)}$$

### 2. 乘法定理

$$\begin{aligned} P(AB) &= P(A)P(A|B) \\ P(AB) &= P(B)P(B|A) \end{aligned}$$

### 3. 事件独立性

相互独立: 设 A, B 为两个事件, 如果  $P(AB) = P(A)P(B)$ , 则称事件 A 与事件 B 相互独立。

## 全概率公式

### 4. 全概率公式

$$P(B) = P(B | A_1)P(A_1) + \cdots + P(B | A_n)P(A_n)$$

## 5. 贝叶斯公式

因果关系  $A \rightarrow B$ , ,  $A_1, A_2, A_3, \dots, A_n$  是一个完备事件组。

$$P(B | A_i) = \frac{P(A_i B)}{P(B)} = \frac{P(A_i)P(B | A_i)}{\sum_{i=1}^n P(A_i)P(B | A_i)}$$

## 事件独立性

事件A、B独立的充要条件是

$$\begin{aligned} P(A | B) &= P(A), P(B) > 0 \\ P(B | A) &= P(B), P(A) > 0 \end{aligned}$$

推倒：

$$P(AB) = P(A)P(B)$$

$$P(AB) = P(A | B)P(B)$$

$$\Rightarrow P(A)P(B) = P(A | B)P(B)$$

$$\Rightarrow P(A) = P(A | B)$$

## 不相容性与独立性

结论：互不相容与相互独立不能同时独立。

$$\text{证明: } A \cap B = \emptyset \implies P(AB) = 0$$

$$P(A) \neq 0, P(B) \neq 0$$

$$P(AB) \neq P(A)P(B)$$

so AB不独立

特例：S和 $\phi$

## 多个事件的独立

### 三个事件的独立

#### 1. ABC 两两独立

$$\begin{cases} P(AB) = P(A)P(B) \\ P(AC) = P(A)P(C) \\ P(BC) = P(B)P(C) \end{cases}$$

#### 2. ABC 相互独立

$$\begin{cases} P(AB) = P(A)P(B) \\ P(AC) = P(A)P(C) \\ P(BC) = P(B)P(C) \\ P(ABC) = P(A)P(B)P(C) \end{cases}$$

## n个事件的独立性

定义 设  $A_1, A_2, \dots, A_n$  为  $n$  个事件, 如果对于任意的  $k (1 < k \leq n)$ , 和任意的  $1 \leq i_1 \leq i_2 \leq \dots \leq i_k \leq n$  有等式

$$P(A_{i_1}A_{i_2}\dots A_{i_k}) = P(A_{i_1})P(A_{i_2})\dots P(A_{i_k})$$

则称  $A_1, A_2, \dots, A_n$  为相互独立的事件.

性质:

(1) 若事件  $A_1, A_2, \dots, A_n (n \geq 2)$  相互独立,

则其中的任意  $k (2 \leq k \leq n)$  个事件也相互独立

(2) 若事件  $A_1, A_2, \dots, A_n (n \geq 2)$  相互独立,

则将  $A_1, A_2, \dots, A_n (n \geq 2)$  中任意多个

事件换成其对立事件, 所得新的  $n$  个事件

仍相互独立

(3) 若  $A_1, A_2, \dots, A_n$  是相互独立的事件, 则

$$\begin{aligned} P(A_1 \cup A_2 \cup \dots \cup A_n) &= 1 - P(\overline{A_1 \cup A_2 \cup \dots \cup A_n}) \\ &= 1 - P(\overline{A_1} \dots \overline{A_n}) = 1 - P(\overline{A_1})P(\overline{A_2}) \dots P(\overline{A_n}) \end{aligned}$$

## 小概率事件

特别的, 如果有  $P(A_1) = P(A_2) = \dots = P(A_n) = p$

$$\text{则有 } P\left(\bigcup_{i=1}^n A_i\right) = 1 - (1-p)^n$$

$$\text{当 } n \rightarrow \infty \text{ 时, } P\left(\bigcup_{i=1}^n A_i\right) = 1 - (1-p)^n \rightarrow 1$$

结论: 小概率事件虽然在一次实验中几乎不可能发生, 但是迟早要发生

## 离散型随机分布

分布与数字特征

概率质量函数: 离散型随机变量

概率密度函数: 连续型随机变量

## 二项分布

$n$  重 Bernoulli 试验中,  $X$  是事件  $A$  在  $n$  次试

验中发生的次数,  $P(A) = p$ , 若

$$P_n(k) = P(X = k) = C_n^k p^k (1 - p)^{n-k}, \quad k = 0, 1, \dots, n$$

则称  $X$  服从参数为  $n, p$  的二项分布, 记作

$$X \sim (n, p)$$

0-1 分布是  $n = 1$  的二项分布

### 二项分布中最可能出现的次数与推倒

若  $P(X = k) \geq P(X = j), j = X$  可取的一切值则称为  $k$  为最有可能出现的次数:

$$p_k = P(X = k) = C_n^k p^k (1 - p)^{n-k}, \quad k = 0, 1, \dots, n$$

$$\left. \begin{aligned} \frac{p_{k-1}}{p_k} &= \frac{(1-p)k}{p(n-k-1)} \leq 1 \\ \frac{p_k}{p_{k+1}} &= \frac{(1-p)(k+1)}{p(n-k)} \geq 1 \end{aligned} \right\}$$

$$\implies (n+1)p - 1 \leq k \leq (n+1)p$$

当  $(n+1)p = Z$  时, 在  $k = (n+1)p$  和  $k = (n+1)p - 1$  处取的最大值

当  $(n+1)p \neq Z$  时, 在  $k = [(n+1)p]$  处的概率取得最大值

## 几何分布

$$X \sim G(p)$$

$$P(X = k) = pq^{k-1}$$

$X$  表示贝努力实验中首次成功事件出现所要进行的试验次数

## Poisson 分布

$$X \sim (\lambda)$$

$$P\{X = k\} = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \dots$$

### 泊松定理

在二项分布  $B(n, p_n)$  中, 如果  $\lim n p_n = \lambda$  则成立:

$$\lim_{n \rightarrow \infty} C_n^k p_n^k (1 - p_n)^{n-k} = \frac{\lambda^k}{k!} e^{-\lambda} \quad (k = 0, 1, \dots)$$

## 连续型随机分布

$$F(x) = P(X \leq x)$$

性质：

1. 单调增
2.  $F(-\infty) = 0, F(+\infty) = 1$
3. 右连续，即  $F(x+0) = F(x)$

常用公式：

$$P(X \leq b) = F(b)$$

$$P(a < X \leq b) = F(b) - F(a)$$

$$P(X > b) = 1 - F(b)$$

$$P(X < b) = F(b-0)$$

## 概率密度

概率密度函数 probability density function PDF

分布函数与概率密度函数的关系

$$F(x) = \int_{-\infty}^x f(t)dt$$

$f(x)$ 就称为概率密度函数

性质：

1.  $f(x) \geq 0$
2.  $\int_{-\infty}^{+\infty} f(x)dx = 1$
3.  $F(x)$ 是连续函数
4. 若  $f(x)$ 在  $x$  处连续，则  $F'(x) = f(x)$
5. 连续型随机变量  $X$  在一个点上的取值概率恒为 0
6.  $P(X \in I) = \int_I f(x)dx, I = (a, b) \text{ or } (a, b], \text{ or } [a, b) \text{ or } [a, b]$

注意：一般的，同一个连续型随机变量  $X$  的概率密度函数可以有很多个，但它们只在有限个点和可数个点的取值不同。所以连续型随机变量  $X$  的概率密度函数“几乎处处”唯一的。

## 计算

1. 分布函数  $F(x)$  是  $f(x)$  的变上限积分函数
2.  $F'(x) = f(x)$
3.  $\int_{-\infty}^{+\infty} f(x)dx = 1$
4.  $P(a \leq X \leq b) = F(b) - F(a) = \int_a^b f(x)dx$
5. 连续型随机变量  $X$  任取一实数的概率值为 0

$$P(X = a) = 0$$

$$P(a < X < b) = P(a < X \leq b) = P(a \leq X \leq b)$$

## 数字特征

### 1. 数学期望

定义:

$$E(X) = x_1 p_1 + x_2 p_2 + \cdots + x_k p_k + \cdots$$

性质:

- (1)  $E(aX + b) = aE(X) + b$
- (2)  $E(aX) = aE(X)$
- (3)  $E(X + b) = E(X) + b$
- (4)  $E(b) = b$
- (5)  $E(X + Y) = E(X) + E(Y)$
- (6)  $E(f(\xi)) = \sum_k f(x_k) P_K$

### 2. 方差

定义

$$D(\xi) = E[\xi - E(\xi)]^2$$

性质

- (1)  $D(c) = 0$
- (2)  $D(k\xi) = k^2 D(\xi)$
- (3)  $D(\xi + b) = D(\xi)$
- (4)  $D(k\xi + b) = k^2 D(\xi)$

## 经典分布的数字特征

分布名称	记号	概率分布	期望	方差
二项分布	$B(n, p)$	$P(X = k) = C_n^k p^k q^{n-k}$	$np$	$npq$
几何分布	$Ge(p)$	$P(X = k) = (1 - p)^{k-1} p$	$1/p$	$q/p^2$
泊松分布	$P(\lambda)$	$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} (k = 0, 1, \dots, \lambda > 0)$	$\lambda$	$\lambda$
均匀分布	$U_{(a,b)}$	$f(x) = \frac{1}{b-a} (a < x < b)$	$(a+b)/2$	$(b-a)^2/12$
指数分布	$\pi(\lambda)$	$f(x) = \begin{cases} \lambda e^{-\lambda}, & x \geq 0 \\ 0, & x < 0 \end{cases}$	$1/\lambda$	$1/\lambda^2$
正态分布	$N(\mu, \delta^2)$	$f(x) = \frac{1}{\sqrt{2\pi}\delta} \exp\left[-\frac{(x-\mu)^2}{2\delta^2}\right]$	$\mu$	$\delta^2$

## 正态分布

定义

标准正态分布

指数分布

均匀分布

## 多元随机变量

---

联合, 边缘, 条件

联合分布列

边缘分布列

条件分布列

## 独立与相关

独立性

条件独立

期望和方差

协方差协方差矩阵

随机变量的相关系数

## 🍑 多元正态分布

常用二维连续型分布

1. 均匀分布
2. 二元正态分布
3. 二次型

## 4. n元正态分布

性质

### 相关系数

相关系数公式

$$R(X, Y) = E \left( \frac{X - E(X)}{\sigma(X)} \cdot \frac{Y - E(Y)}{\sigma(Y)} \right)$$

二维正态分布中X和Y的相关系数  $R(X, Y)$

$$R(X, Y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-r^2}} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \frac{(x - \mu_x)}{\sigma_x} \cdot \frac{(y - \mu_y)}{\sigma_y} e^{-u(x,y)} dx dy$$
$$u(x, y) = \frac{(x - \mu_x)^2}{2\sigma_x^2} + \frac{1}{2(1-r^2)} \left[ \frac{(y - \mu_y)}{\sigma_y} + \frac{r(x - \mu_x)}{\sigma_x} \right]^2$$

### 统计量

$g(X_1, X_2, X_3, \dots, X_n)$  不含未知参数

#### 常用统计量

1. 样本均值:  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$

观察值:  $\bar{x} = \frac{1}{n} \sum_{i=1}^n X_i$

2. 样本方差

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

3. 样本标准差

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

观察值:  $X \rightarrow x$

4. 样本k阶 (原点) 矩

$$A_k = \frac{1}{n} \sum_{i=1}^n X_i^k, k = 1, 2, \dots$$

观察值:  $X \rightarrow x$

5. 样本k阶中心距

$$B_k = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^k, k = 2, 3, \dots$$

## 抽样分布

### 卡方分布

设 $X_1, X_2, \dots, X_n$ 相互独立, 且都服从标准正态分布 $N(0, 1)$ , 则

$$\sum_{i=1}^n X_i^2 \sim \chi^2(n)$$

自由度为 $n$ 的 $\chi^2(n)$ 的密度函数为:

$$f(x) = \begin{cases} \frac{1}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})} e^{-\frac{x}{2}} x^{n-1}, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

其中,

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

在 $x>0$ 的时候收敛, 称为 $\chi$ 函数, 具有性质

$$\begin{aligned} \Gamma(x+1) &= x\Gamma(x) \\ \Gamma(1) &= 1, \Gamma(1/2) = \sqrt{\pi} \\ \Gamma(n+1) &= n! \quad (n \in N) \end{aligned}$$

### 卡方分布的性质

1.  $E(\chi^2(n)) = n, D(\chi^2(n)) = 2n$
2. 若 $X_1 = \chi^2(n_1), X_2 = \chi^2(n_2), X_1, X_2$ 相互独立, 则

$$X_1 + X_2 \sim \chi^2(n_1 + n_2)$$

3.  $n \rightarrow \infty$  正态分布
4. 上 $\alpha$ 分位点

$$P\{X \geq \chi_\alpha^2(n)\} = \alpha$$

则称 $\chi_\alpha^2(n)$ 为 $\chi^2(n)$ 分布的上 $\alpha$ 分位点

## t分布

设 $X \sim N(0, 1), Y \sim \chi^2(n), X, Y$ 相互独立,

$$T = \frac{X}{\sqrt{Y/n}}$$

则 $T$ 所服从的分布称为自由度为 $n$ 的t分布其密度函数为

$$f(t) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}} \quad -\infty < t < \infty$$

## 性质

1.  $f_n(t)$ 是偶函数

$$n \rightarrow \infty, f_n(t) \rightarrow \varphi(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$$

2. t分布的上 $\alpha$ 分位数 $t_\alpha$ 与双侧的 $\alpha$ 分位数 $t_{\alpha/2}$ 有表可查

3.  $P(T > t_\alpha) = \alpha, -t_\alpha = t_{1-\alpha}$

4.  $P(T > t_{\alpha/2}) = \frac{\alpha}{2}, P(T > t_{\alpha/2}) = \alpha$

## F分布

设 $X \sim \chi^2(n), Y \sim \chi^2(m)$ , XY相互独立, 令

$$F = \frac{X/n}{Y/m}$$

则F所服从的分布称为第一自由度为n,第二自由度为m的F分布,其密度函数为

$$f(t, n, m) = \begin{cases} \frac{\Gamma(\frac{n+m}{2})}{\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})} \left(\frac{n}{m}\right)^{\frac{n}{2}} t^{\frac{n}{2}-1} \left(1 + \frac{n}{m}t\right)^{-\frac{n+m}{2}}, & t > 0 \\ 0, & t \leq 0 \end{cases}$$

## 性质

1.  $F \sim F(n, m)$ , 则 $\frac{1}{F} \sim F(m, n)$
2.  $F(n, m)$ 的上 $\alpha$ 分位数 $F_\alpha(n, m)$ 有表可查:  $P(F > F_\alpha(n, m)) = \alpha$
3. 分位点性质

$$\begin{aligned} F_{1-\alpha}(n, m) &= \frac{1}{F_\alpha(m, n)} \\ [t_{1-\frac{\alpha}{2}}(n)]^2 &= F_\alpha(1, n) \end{aligned}$$

## 其他分布

Gamma分布

Beta分布

Fisher Z分布

指数结构

## 推断分布

# 大数定律

由样本推断总体的依据

## 切比雪夫大数定律

## 中心极限定理

中心极限定理:设从均值为 $\mu$ , 方差为 $\sigma^2$ 的一个任意总体中抽取容量为n的样本, 当n充分大时, 样本均值的抽样分布近似服从均值为 $\mu$ 、方差为 $\frac{\sigma^2}{n}$ 的正态分布。

# ch4参数估计

基本思想: 用样本值取估计参数的取值

## 参数的点估计

### 矩估计

用样本矩来代替总体矩,从而得到总体分布参数 的一种估计量.这种估计方法称为矩估计法

样本k阶 (原点) 矩

$$A_k = \frac{1}{n} \sum_{i=1}^n X_i^k, k = 1, 2, \dots$$

观察值:  $X \rightarrow x$

## 极大似然法

### 极大似然函数

设总体X的分布类型已知, 但是含有参数 $\theta$

设离散型总体X的概率分布为 $p(x, \theta)$ ,则样本 $(X_1, X_2, \dots, X_n)$ 的联合概率密度函数称为似然函数

$$L(\theta) = f(x_1; \theta) f(x_2; \theta) \cdots f(x_n; \theta) = \prod_{i=1}^n f(x_i; \theta)$$

### 极大似然参数估计值

若 $L(\theta)$  在 $\hat{\theta} = \hat{\theta}(x_1, x_2, \dots, x_n)$  处取到极大值, 则称 $\hat{\theta}(x_1, x_2, \dots, x_n)$  为 $\theta$  的极大似然估计值。

## 求解步骤

1. 求似然函数  $L(\theta)$
2. 求出  $\ln L(\theta)$  及方程  $\frac{d}{d\theta} \ln L(\theta) = 0$
3. 解上述方程得到极大似然估计值

$$\hat{\theta} = \hat{\theta}(x_1, x_2, \dots, x_n)$$

4. 解3方程得到极大似然估计量

$$\hat{\theta} = \hat{\theta}(X_1, X_2, \dots, X_n)$$

## 估计量的评选标准

常用  
标准  $\begin{cases} (1) \text{ 无偏性 (Unbiased Estimator)} \\ (2) \text{ 有效性} \\ (3) \text{ 一致性 (consistency)} \end{cases}$

### 无偏估计

参数等于均值:  $E[\hat{\theta}(X_1, X_2, \dots, X_n)] = \theta$

例题:

$$\begin{aligned} E(S^2) &= E\left[\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2\right] \\ &= E\left[\frac{1}{n-1} \left( \sum_{i=1}^n X_i^2 - n\bar{X}^2 \right)\right] \\ &= \frac{1}{n-1} \left[ \sum_{i=1}^n E(X_i^2) - nE(\bar{X}^2) \right] \\ &= \frac{1}{n-1} \left[ \sum_{i=1}^n (\sigma^2 + \mu^2) - n \left( \frac{\sigma^2}{n} + \mu^2 \right) \right] = \sigma^2 \end{aligned}$$

### 有效性

方差更小的更有效:  $D(\hat{\theta}_1) \leq D(\hat{\theta}_2)$

$$\hat{\theta}_1 = \hat{\theta}_1(X_1, X_2, \dots, X_n)$$

$$\hat{\theta}_2 = \theta_2(X_1, X_2, \dots, X_n)$$

## 一致性

一致估计量的意义在于：只要样本容量足够大，就可以使一致估计量与参数真实值之间的差异大于  $\varepsilon$  的概率足够地小，也就是估计量可以用任意接近于 1 的概率把参数真实值估计到任意的精度。

这种性质是针对样本容量  $b \rightarrow +\infty$  而言，对于一个固定的样本容量  $n$ ，一致性是无意义的。

## 区间估计

存在两个统计量  $\underline{\theta}$ , 使得  $\bar{\theta}$

$$P \left\{ \underline{\theta}(X_1, X_2, \dots, X_n) < \theta < \bar{\theta}(X_1, X_2, \dots, X_n) \right\} = 1 - \alpha$$

则称区间  $(\underline{\theta}, \bar{\theta})$  为参数  $\theta$  的置信度为  $1 - \alpha$  的置信区间

置信区间

1. 精度：区间长度
2. 置信度  $1 - \alpha$

## 求置信区间的步骤

1. 构造一个含有未知参数而不含有其他参数的样本函数（随机变量）

$$W = W(X_1, X_2, \dots, X_n : \theta)$$

2. 根据  $W = W(X_1, X_2, \dots, X_n : \theta)$  的分布类型（自由度）及给定置信度  $1 - \alpha$ ，查  $\alpha$  或  $\alpha/2$  或  $1 - \alpha/2$  分位点定出分位数  $a$  和  $b$ ，使得

$$P \{ a < W(X_1, X_2, \dots, X_n) < b \} = 1 - \alpha \quad \text{方差已知时, 均值的区间估计}$$

3. 解不等式

$$a < W(X_1, X_2, \dots, X_n; \theta) < b$$

这时必有

$$P \left\{ \hat{\theta}_1(X_1, X_2, \dots, X_n) < \theta < \hat{\theta}_2(X_1, X_2, \dots, X_n) \right\} = 1 - \alpha$$

于是  $(\hat{\theta}_1, \hat{\theta}_2)$  即为  $\theta$  的置信度为  $1 - \alpha$  的置信区间。

## 均值的区间估计

## 一致最小方差无偏估计

### 最小均方误差准则

$$MSE_{\theta}(\hat{\theta}) = E(\hat{\theta} - \theta)^2$$

如果  $MSE_{\theta}(\hat{\theta}) < +\infty$

$$\begin{aligned} MSE_{\theta}(\hat{\theta}) &= \text{Var}_{\theta}(\hat{\theta}) + b^2(\theta, \hat{\theta}) \\ b(\theta, \hat{\theta}) &= E_{\theta}(\hat{\theta} - \theta) \end{aligned}$$

$$E[g^*(\tilde{X}) - g(\theta)]^2 \leq E[\hat{g}(\tilde{X}) - g(\theta)]^2$$

### 一致最小方差无偏估计(UMVUE)

是在无偏估计类中，使均方误差达到最小的估计量

## Cramer-Rao公式

### CR正则分布族

单参数密度函数满足以下五个条件为CR正则分布族

1. 参数空间是直线上的某个开区间
2. 导数存在
3.  $p(x, \theta)$ 不依赖于参数
4. 对概率密度函数p的积分与微分运算可以交换
5. 下列数学期望存在

$$0 < I(\theta) = E_{\theta} \left\{ \frac{\partial}{\partial \theta} \ln p(X; \theta) \right\}^2 < +\infty$$

### C-R不等式

定理：正则分布族无偏估计的下界，也称作C-R下界

$$D_{\theta}[\hat{g}(\tilde{X})] \geq \frac{[g'(\theta)]^2}{nI(\theta)}, \theta \in \Theta$$

证明：

$$\begin{aligned}
\frac{\partial}{\partial \theta} \ln p(x_1, \dots, x_n; \theta) &= \sum_{i=1}^n \frac{\partial}{\partial \theta} \ln p(x_i; \theta) \\
S(\tilde{X}; \theta) &= \frac{\partial}{\partial \theta} \ln p(X_1, \dots, X_n; \theta) \\
E_\theta \left\{ \frac{\partial}{\partial \theta} \ln p(X_i; \theta) \right\} &= \int \frac{\partial}{\partial \theta} \ln p(x_i, \theta) p(x_i, \theta) dx_i \\
&= \int \frac{\partial}{\partial \theta} p(x_i, \theta) dx_i = \frac{d}{d\theta} \int p(x_i, \theta) dx_i = \frac{d}{d\theta} 1 = 0 \\
E_\theta \{ S(\tilde{X}, \theta) \} &= \sum_{i=1}^n E_\theta \left\{ \frac{\partial}{\partial \theta} \ln p(X_i, \theta) \right\} \\
D_\theta \{ S(\tilde{X}, \theta) \} &= D_\theta \left\{ \sum_{i=1}^n \frac{\partial}{\partial \theta} \ln p(X_i, \theta) \right\} \\
&= \sum_{i=1}^n D_\theta \left\{ \frac{\partial}{\partial \theta} \ln p(X_i, \theta) \right\} \\
&= \sum_{i=1}^n E_\theta \left\{ \frac{\partial}{\partial \theta} \ln p(X_i, \theta) \right\}^2 = nI(\theta)
\end{aligned}$$

可以看到C-R不等式的右端与参数g(θ)的变化率的平方成正比, 与总体所在分布族的Fisher信息量的n倍成反比.

## 有效估计

无偏估计的效率:

$$e_n = \frac{[g'(\theta)]^2 / nI(\theta)}{D_\theta(\hat{g}(\tilde{X}))}$$

$e_n = 1$  有效无偏估计

$\lim_{n \rightarrow \infty} e_n = 1$  漐进有效(无偏)估计

结论:

有效估计一定是UMVUE, 但很多 UMVUE 不是有效估计, 这是因为 C-R 下界偏小, 在很多场合达不到.

等式成立的充要条件:

$$\begin{aligned}
S(\tilde{X}, \theta) - ES(\tilde{X}, \theta) &= t(\hat{g}(\tilde{X}) - g(\theta)) \\
I(\theta) &= -E \left[ \frac{\partial^2}{\partial \theta^2} \ln f(x, \theta) \right]
\end{aligned}$$

推论:

1. 条件: 求s, 可表示为

$$\begin{aligned}
S(\tilde{X}, \theta) &= \frac{\partial}{\partial \theta} \ln f(x_1, x_2, \dots, x_n, \theta) = c(\theta)(\hat{g}(\tilde{X}) - g(\theta)) \\
E(\hat{g}(\tilde{X})) &= g(\theta)
\end{aligned}$$

2. 若上式成立

$$I(\theta) = \frac{c(\theta)g'(\theta)}{n}$$

判断方法:

1. 求I 主定理
2. 求S 用推论

## 例题

1.  $X \sim B(1, p)$ , 求  $p$  的有效估计量
2.  $X \sim \pi(1, \lambda)$ , 求  $\lambda$  的有效估计量
3.  $X \sim N(\mu, \sigma^2)$ , 求参数的有效估计量

# ch5 贝叶斯估计

## 贝叶斯方法

1. 选择先验分布对参数的信念
2. 在给定参数情况下对  $x$  的信念
3. 得到数据后更新我们的信念, 计算后验分布
4. 从后验分布中得到点估计和区间估计

## 贝叶斯公式

贝叶斯推理就是在不完全情报下, 对部分未知的状态用主观概率估计, 然后用贝叶斯公式对先验概率进行修正, 最后再利用修正概率做出最优决策。

贝叶斯决策理论方法是统计决策中的一个基本方法, 其基本思想是:

- 1、已知条件概率密度参数表达式和先验概率。
- 2、利用贝叶斯公式转换成后验概率。
- 3、根据后验概率大小进行决策分

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{i=1}^n P(B|A_i)P(A_i)}$$

其中  $\sum_{i=1}^n P(A_i) = 1 \quad \sum_{i=1}^n P(B|A_i)P(A_i) = P(B)$

## 先验分布: $\pi(\theta)$

$\pi(\theta)$ : 对未知参数的先验信息用一个分布形式来表示, 此分布称为未知参数的先验分布.

## 后验分布: $h(\theta | x)$

在抽取样本之前, 人们对未知参数有个了解, 即先验分布。抽取样本之后, 由于样本中包含未知参数的信息, 而这些关于未知参数新的信息可以帮助人们修正抽样之前的先验信息

$$q(x_1, x_2, \dots, x_n) = \prod_{i=1}^n p(x_i, \theta)$$

而样本值是在知道参数 $\theta$ 的先验分布的前提下得到的，因而上述分布可以改写为

$$q(x | \theta) = q(x_1, x_2, \dots, x_n | \theta) = \prod_{i=1}^n p(x_i | \theta)$$

又由于参数 $\theta$ 和样本 $x$ 的联合分布可以表示为

$$\begin{aligned} f(x, \theta) &= q(x | \theta)\pi(\theta) = m(x)h(\theta | x) \\ \implies h(\theta | x) &= \frac{q(x | \theta)\pi(\theta)}{m(x)}, \quad \left( m(x) = \int_{\Theta} q(x | \theta)\pi(\theta) d\theta \right) \end{aligned}$$

可以根据数据量的增加一直修正参数

必考题：

为了提高某产品的质量，公司经理考虑增加投资来改进生产设备，预计需投资90万元，但从投资效果来看，顾问们提出两种不同的意见：

后验分布更能准确描述事情真相

## 共轭先验分布

在贝叶斯统计中，如果后验分布与先验分布属于同类，则先验分布与后验分布被称为共轭分布，而先验分布被称为似然函数的共轭先验。

样本X的分布为二项分布 $b(n, \theta)$ 时，假如 $\theta$ 的先验分布为 $\beta$ 分布，则用贝叶斯估计算得的后验分布仍然是 $\beta$ 分布，只是其中的参数不同。这样的先验分布( $\beta$ 分布)称为参数 $\theta$ 的共轭先验分布。

## 贝叶斯估计

- 使后验密度 $\pi(\theta|x)$ 达到最大的值 $\theta_{MD}$ 称为最大后验估计；
- 后验分布的中位数 $\hat{\theta}_{Me}$ 称为后验中位数估计；
- 后验分布的期望值 $\hat{\theta}_E$ 称为 $\theta$ 的后验期望值估计，这三个估计都称为贝叶斯估计，记为 $\hat{\theta}_B$ 。

必考题：

设一批产品的不合格率为 $\theta$ ，检查是一个一个进行，直到发现第一个不合格品为止，若 $X$ 为发现第一个不合格品时已检查的产品数，则 $X$ 服从几何分布，其分布列为

## ch6假设检验

### 假设检验基本思想

## 假设检验步骤

1. 根据实际问题建立原假设和备择假设
2. 选择合适的统计量
3. 给定显著的水平和 $\alpha$ 确定的临界值
4. 根据样本的值计算出统计量的数值并作出决策

### 建立假设

原假设：一般研究者想收集证据予以反对的假设

备择假设：一般研究者想收集证据予以支持的假设

三种形式：

1. 双侧检验  $H_0 : \mu = \mu_0, H_1 : \mu \neq \mu_0$
2. 左侧检验  $H_0 : \mu \geq \mu_0, H_1 : \mu < \mu_0$
3. 右侧检验  $H_0 : \mu \leq \mu_0, H_1 : \mu > \mu_0$

### 设计检验统计量步骤

1. 根据样本观测结果计算得到的，并据以对原假设和备择假设作出决策的某个样本统计量
2. 标准化检验统计量

$$\text{标准化检验统计量} = \frac{\text{点估计量}-\text{假设值}}{\text{点估计量的抽样标准差}}$$

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \quad t_{(n-1)} = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

3. 拒绝域和接受域的确定

### 两类错误

$$\begin{aligned}\alpha &= P\{\text{第I类错误}\} = P\{\text{拒绝 } H_0 \mid H_0 \text{ 是真实的}\}, \\ \beta &= P\{\text{第II类错误}\} = P\{\text{接受 } H_0 \mid H_0 \text{ 是错误的}\}.\end{aligned}$$

## 总体均值的检验

### 一个总体均值的检验

# 总体均值的检验规则 (正态, 小样本, 方差已知)

假设	双侧检验	左侧检验	右侧检验
假设形式	$H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$	$H_0: \mu = \mu_0$ $H_1: \mu < \mu_0$	$H_0: \mu = \mu_0$ $H_1: \mu > \mu_0$
统计量	$\sigma$ 已知	$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$	
拒绝域	$ z  > z_{\alpha/2}$	$z < -z_\alpha$	$z > z_\alpha$
P值决策	$P < \alpha$	拒绝 $H_0$	

# 总体均值的检验规则 (正态, 方差未知, 小样本情形)

假设	双侧检验	左侧检验	右侧检验
假设形式	$H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$	$H_0: \mu \geq \mu_0$ $H_1: \mu < \mu_0$	$H_0: \mu \leq \mu_0$ $H_1: \mu > \mu_0$
统计量	总体 $\sigma$ 未知	$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$	
拒绝域	$ t  > t_{\alpha/2}(n-1)$	$t < -t_\alpha(n-1)$	$t > t_\alpha(n-1)$
P值决策	$P < \alpha$	拒绝 $H_0$	

大样本

假设	双侧检验	左侧检验	右侧检验
假设形式	$H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$	$H_0: \mu \geq \mu_0$ $H_1: \mu < \mu_0$	$H_0: \mu \leq \mu_0$ $H_1: \mu > \mu_0$
统计量	$\sigma$ 已知:	$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$	
	$\sigma$ 未知:	$z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$	
拒绝域	$ z  > z_{\alpha/2}$	$z < -z_\alpha$	$z > z_\alpha$
P值决策	$P < \alpha$	拒绝 $H_0$	

## 两个总体均值之差的检验

小样本情形

假设	双侧检验	左侧检验	右侧检验
假设形式	$H_0: \mu_1 - \mu_2 = 0$ $H_1: \mu_1 - \mu_2 \neq 0$	$H_0: \mu_1 - \mu_2 \geq 0$ $H_1: \mu_1 - \mu_2 < 0$	$H_0: \mu_1 - \mu_2 \leq 0$ $H_1: \mu_1 - \mu_2 > 0$
统计量	总体 $\sigma$ 未知	$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$	
拒绝域	$ t  > t_{\alpha/2} (n_1 + n_2 - 2)$	$t < -t_\alpha$	$t > t_\alpha$
P值决策	$P < \alpha$	拒绝 $H_0$	

大样本  $n_1 > 30$  和  $n_2 > 30$

假设	双侧检验	左侧检验	右侧检验
假设形式	$H_0: \mu_1 - \mu_2 = 0$ $H_1: \mu_1 - \mu_2 \neq 0$	$H_0: \mu_1 - \mu_2 \geq 0$ $H_1: \mu_1 - \mu_2 < 0$	$H_0: \mu_1 - \mu_2 \leq 0$ $H_1: \mu_1 - \mu_2 > 0$
统计量	$\sigma_1^2, \sigma_2^2$ 已知	$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$	
	$\sigma_1^2, \sigma_2^2$ 未知	$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$	
拒绝域	$ z  > z_{\alpha/2}$	$z < -z_\alpha$	$z > z_\alpha$
P值决策	$P < \alpha$	拒绝 $H_0$	

## 总体比率的检验

假设	双侧检验	左侧检验	右侧检验
假设形式	$H_0: \pi = \pi_0$ $H_1: \pi \neq \pi_0$	$H_0: \pi \geq \pi_0$ $H_1: \pi < \pi_0$	$H_0: \pi \leq \pi_0$ $H_1: \pi > \pi_0$
统计量	$z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1 - \pi_0)}{n}}}$		
拒绝域	$ z  > z_{\alpha/2}$	$z < -z_\alpha$	$z > z_\alpha$
P值决策	$P < \alpha$ 拒绝 $H_0$		

## 两个总体比率之差的检验

假设	双侧检验	左侧检验	右侧检验
假设形式	$H_0: \pi_1 - \pi_2 = 0$ $H_1: \pi_1 - \pi_2 \neq 0$	$H_0: \pi_1 - \pi_2 \geq 0$ $H_1: \pi_1 - \pi_2 < 0$	$H_0: \pi_1 - \pi_2 \leq 0$ $H_1: \pi_1 - \pi_2 > 0$
统计量	$z = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$		
拒绝域	$ z  > z_{\alpha/2}$	$z < -z_\alpha$	$z > z_\alpha$
P值决策	$P < \alpha$ 拒绝 $H_0$		

## 总体方差的检验

一个样本与总体方差的比较- 卡方检验

假设	双侧检验	左侧检验	右侧检验
假设形式	$H_0: \sigma^2 = \sigma_0^2$ $H_1: \sigma^2 \neq \sigma_0^2$	$H_0: \sigma^2 \geq \sigma_0^2$ $H_1: \sigma^2 < \sigma_0^2$	$H_0: \sigma^2 \leq \sigma_0^2$ $H_1: \sigma^2 > \sigma_0^2$
统计量	$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$		
拒绝域	$\chi^2 > \chi_{\alpha/2}^2(n-1)$ $\chi^2 < \chi_{1-\alpha/2}^2(n-1)$	$\chi^2 < \chi_{\alpha/2}^2(n-1)$	$\chi^2 > \chi_{\alpha/2}^2(n-1)$
$P$ 值决策	$P < \alpha$ 拒绝 $H_0$		

两个样本方差的比较- F检验

假设	双侧检验	左侧检验	右侧检验
假设形式	$H_0: \sigma_1^2/\sigma_2^2 = 1$ $H_1: \sigma_1^2/\sigma_2^2 \neq 1$	$H_0: \sigma_1^2/\sigma_2^2 \geq 1$ $H_1: \sigma_1^2/\sigma_2^2 < 1$	$H_0: \sigma_1^2/\sigma_2^2 \leq 1$ $H_1: \sigma_1^2/\sigma_2^2 > 1$
统计量	$F = \frac{s_1^2}{s_2^2}$ 或 $F = \frac{s_2^2}{s_1^2}$		
拒绝域	$F > F_{\alpha/2, n_2-1}$		

## 检验的p值

### 假设检验的功效函数

将两类错误概率用统一的函数表示出来

## ch7 非参数检验

### 卡方拟合优度检验

# 不含未知参数

## 检验步骤

1. 将总体划分为r个子集：理论频数 $npi>=5$ , 最好在10以上, 否则要合并区间
2. 构造假设：并在 $H_0$ 为真的前提下，计算 $p_{i0}$ , 然后求得理论频数 $npi_0$
3. 统计事件 $A_i$ 的实际频数 $n_i$
4. 构造检验统计量计算结果：

$$\chi_n^2 = \sum_{i=1}^m \frac{(n_i - np_{i0})^2}{np_{i0}}$$

5. 拒绝域：

$$W = \{\chi^2 \geq \chi_\alpha^2(m-1)\}$$

总体离散和总体连续的区别

# 含未知参数

## 检验步骤

1. 利用MLE估计r个参数
2. 求 $p_i$ 的估计 $\hat{p}_i$
3. 计算统计量

$$\chi_n^2 = \sum_{i=1}^m \frac{(n_i - n\hat{p}_{i0})^2}{n\hat{p}_{i0}}$$

拒绝域：r为参数个数

$$W = \{\chi^2 \geq \chi_\alpha^2(m-r-1)\}$$

# 列联表的独立检验

列联表：

		$X_2$				合计
		$B_1$	$B_2$	...	$B_c$	
$X_1$	$A_1$	$n_{11}$	$n_{12}$	...	$n_{1c}$	$n_{1\bullet}$
	$A_2$	$n_{21}$	$n_{22}$	...	$n_{2c}$	$n_{2\bullet}$
	$\vdots$			...	$\vdots$	$\vdots$
	$A_r$	$n_{r1}$	$n_{r2}$	...	$n_{rc}$	$n_{r\bullet}$
合计		$n_{\bullet 1}$	$n_{\bullet 2}$		$n_{\bullet c}$	$n$

构建理论频数列联表

$$e_{ij} = \frac{n_{i\cdot} * n_{\cdot j}}{n}$$

假设：

$$H_0 : p_{ij} = p_i \cdot p_{\cdot j}, H1 : \text{至少一对不符合}$$

统计量：

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(n_{ij} - n\hat{p}_{i\cdot}\hat{p}_{\cdot j})^2}{n\hat{p}_{i\cdot}\hat{p}_{\cdot j}}$$

拒绝域：

$$W = \{\chi^2 \geq \chi^2_\alpha((r-1)(c-1))\}$$

## 正态性检验

---

判断总体的分布是否为正态分布

### W检验

1. 使用条件：样本  $3 < n < 50$
2. 检验统计量--w统计量

$$W = \frac{\left[ \sum_{i=1}^n (a_i - \bar{a}) (X_{(i)} - \bar{X}) \right]^2}{\sum_{i=1}^n (a_i - \bar{a})^2 \sum_{i=1}^n (X_i - \bar{X})^2}$$

$$W' = \frac{\left[ \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} a_i (X_{(n+1-i)} - X_{(i)}) \right]^2}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

3. 拒绝域

$$\{W \leq W_\alpha\}$$

### D检验

0. 使用条件：样本  $n > 50$
1. 将样本值按照升序排列
2. 计算统计量

$$Y = \frac{(D - 0.28209479)\sqrt{n}}{0.02998598}$$

其中

$$D = \frac{\sum_{i=1}^n \left(i - \frac{n+1}{2}\right) X_{(i)}}{n^{\frac{3}{2}} \sqrt{\sum_{i=1}^n (X_i - \bar{X})^2}}$$

3. 选定显著性水平查询D检验法临界值表得到

$$Y_{\frac{\alpha}{2}}, Y_{1-\frac{\alpha}{2}}$$

4. 得出结论

若  $Y_{\frac{\alpha}{2}} \leq Y \leq Y_{1-\frac{\alpha}{2}}$  则符合正态分布, 否则不符合正态分布

## ch8 方差分析

### 单因素方差分析

$$SSA = \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{x}_i - \bar{x})^2 = \sum_{i=1}^k n_i (\bar{x}_i - \bar{x})^2$$

$$SSE = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$$

$$SST = \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2$$

### 双因素方差分析

## ch9 回归分析

### 一元线性回归的参数估计

#### 模型

1. 模型

$$Y = a + bx + \varepsilon$$
$$\varepsilon \sim N(0, \sigma^2)$$

#### 参数估计-最小二乘法

$$Q(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x)^2$$

1. 目标函数:

$$Q(\hat{\beta}_0, \hat{\beta}_1) = \min_{(\beta_0, \beta_1)} Q(\beta_0, \beta_1)$$

$$\begin{cases} \frac{\partial Q}{\partial \beta_0} \Big|_{\beta_0=\hat{\beta}_0} = 0 \\ \frac{\partial Q}{\partial \beta_1} \Big|_{\beta_1=\hat{\beta}_1} = 0 \end{cases}$$

$$\begin{cases} n\hat{\beta}_0 + \sum_i x_i \hat{\beta}_1 = \sum_i y_i \\ \sum_i x_i \hat{\beta}_0 + \hat{\beta}_1 \sum_i x_i^2 = \sum_i x_i y_i \end{cases}$$

$$\hat{\beta}_1 = \frac{\sum_i x_i y_i - n\bar{x}\bar{y}}{\sum_i x_i^2 - n\bar{x}^2}$$

我们定义：

$$l_{xy} = \sum_i (x_i - \bar{x})(y_i - \bar{y})$$

$$l_{xx} = \sum_i (x_i - \bar{x})^2$$

简化计算1：

$$L_{xx} = \sum_{i=1}^n x_i^2 - \frac{1}{n} \left( \sum_{i=1}^n x_i \right)^2$$

$$L_{xy} = \sum_{i=1}^n x_i y_i - \frac{1}{n} \left( \sum_{i=1}^n x_i \right) \left( \sum_{i=1}^n y_i \right)$$

简化计算2：

$$L_{xx} = \sum_{i=1}^n x_i^2 - n\bar{x}^2$$

$$L_{xy} = \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}$$

从而可以得到：

$$\begin{cases} \hat{\beta}_1 = \frac{l_{xy}}{l_{xx}} \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \end{cases}$$

## 2. 估计量的分布

1.  $\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{l_{xx}}\right)$
2.  $\hat{\beta}_0 \sim N\left(\beta_0, \left(\frac{1}{n} + \frac{\bar{x}^2}{l_{xx}}\right)\sigma^2\right)$

## 3. 估计量的协方差

$$\text{cov}(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\bar{x}}{l_{xx}}\sigma^2$$

## 4. y的分布

$$\hat{y} \sim N \left( \beta_0 + \beta_1 x, \left[ \frac{1}{n} + \frac{(x - \bar{x})^2}{l_{xx}} \right] \sigma^2 \right)$$

## 回归方程的显著性检验

三种方法彼此是等价的（使用时看哪一种方法计算量最少就用哪一个）

### 检验假设

#### 1. 构建原假设

$$H_0 : \beta_1 = 0$$

## 1. F检验

#### 1. 构造统计量

$$S_E = \sum_i (y_i - \hat{y}_i)^2$$

$$S_E/(n-2) \sim \chi^2(n-2)$$

$$\frac{\frac{S_R}{\sigma^2}}{\frac{S_E}{\sigma^2(n-2)}} = \frac{S_R}{S_E/(n-2)} = F \sim F(1, n-2)$$

#### 2. 拒绝域: $W = F > F_\alpha(1, n-2)$

## 2. T检验

#### 1. 构造统计量:

$$t = \frac{\hat{\beta}_1}{\hat{\sigma}/\sqrt{l_{xx}}} = \frac{\frac{\hat{\beta}_1}{\sigma/\sqrt{l_{xx}}}}{\sqrt{\frac{S_E/(n-2)}{\sigma^2/(n-2)}}} \sim t(n-2)$$

#### 2. 拒绝域: $W = |t| > t_{\alpha/2}(n-2)$

## 3. 相关系数的检验

#### 1. 相关系数:

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} = \frac{l_{xy}}{\sqrt{l_{xx}l_{yy}}}$$

检验假设:

$$H_0 : \beta_1 = 0$$

构造统计量：

$$r = \frac{l_{xy}}{\sqrt{l_{xx}l_{yy}}} = \frac{l_{xy}}{l_{xx}} \cdot \sqrt{\frac{l_{xx}}{l_{yy}}} = \hat{\beta}_1 \sqrt{\frac{l_{xx}}{l_{yy}}}$$

拒绝域：

$$\{|r| \geq r_{\frac{\alpha}{2}}(n-2)\}$$

## 预测与控制

### 多元线性回归

## ch10 试验设计

### 试验设计概要

#### 名次解释

1. 指标：用于衡量试验结果好坏的特性值
2. 因子：影响试验结果的因素
3. 水平：因子所处的状态
4. 试验误差：测量值和实际值之间的偏差
5. 试验设计

## 正交化实验设计

三因素、三水平全面试验方案--正交表

正交表需要满足的条件：

1. 每列中含不同数字的个数相同
2. 任一列中的同一数字的那些数字在其他列被不同数字占据，且个数相同。

### 1. 等水平正交表

$L_n(r^m)$

L：正价表记号

n：正交表横行数---试验次数

r：因素水平个数

m：正交纵列数

最低的试验次数 (行数)= $\Sigma$  (每列水平数-1)+1

$$n = m(r - 1) + 1$$

## 2. 混合水平正交表

$$\mathbf{L}_n (r_1^{m_1} \times r_2^{m_2})$$

$$n = m_1(r_1 - 1) + m_2(r_2 - 1) + 1$$

eg.  $\mathbf{L}_8 (4^1 \times 2^4)$

14表示：第一列有4个水平

24:后面4列每列2个水平

## 直观分析法--极差分析法

不考虑交互作用

极差分析法

$$R = \max(K_i) - \min(K_i)$$

方差分析

1. 偏差平方和分解

$$SS_T = SS_{\text{因素}} + SS_{\text{空列(误差)}}$$

2. 自由度分解

$$df_T = df_{\text{因素}} + df_{\text{空列(误列)}}$$

3. 方差

$$MS_{\text{因素}} = \frac{SS_{\text{因素}}}{df_{\text{因素}}}, MS_{\text{误差}} = \frac{SS_{\text{误差}}}{df_{\text{误差}}}$$

4. 构造F统计量

$$F_{\text{因素}} = \frac{MS_{\text{因素}}}{MS_{\text{误差}}}$$

5. 列方差分析表，作F检验

6. 做出判断

若计算出的F值  $F_0 > F_a$ , 则拒绝原假设, 认为该因素或交互作用对试验结果有显著影响; 若  $F_0 \leq F_a$ , 则认为该因素或交互作用对试验结果无显著影响。

总偏差平方和

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

列偏差平方和

$$SS_j = \frac{1}{r} \sum_{i=1}^m K_{ij}^2 - \frac{(\sum_{i=1}^n x_i)^2}{n} (j = 1, 2, \dots, k)$$

## 公式

### 正态分布计算

- (1)  $F(\mu) = 0.5$
- (2)  $F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$
- (3)  $P\{a \leq X < b\} = F(b) - F(a)$   
 $= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$
- (4)  $P\{|X - \mu| < c\} = 2\Phi\left(\frac{c}{\sigma}\right) - 1$

### 协方差

$$\hat{n}_i = n * \hat{p}_i$$

$$\text{Cov}(X, Y) = E[(X - EX)(Y - EY)] = E(XY) - E(X) \cdot E(Y)$$

1. 对称性:  $\text{Cov}(\mathbf{X}, \mathbf{Y}) = \text{Cov}(\mathbf{Y}, \mathbf{X})$
2. 线性性:  $\text{Cov}(a\mathbf{X}, \mathbf{Y}) = a \cdot \text{Cov}(\mathbf{X}, \mathbf{Y})$   
 $\text{Cov}(\mathbf{X}_1 + \mathbf{X}_2, \mathbf{Y}) = \text{Cov}(\mathbf{X}_1, \mathbf{Y}) + \text{Cov}(\mathbf{X}_2, \mathbf{Y})$
3. 若  $X$  和  $Y$  相互独立, 则  $\Rightarrow \text{Cov}(\mathbf{X}, \mathbf{Y}) = \mathbf{0}$   
 因为  $X$  和  $Y$  相互独立  $\Rightarrow E(\mathbf{X}) \cdot E(\mathbf{Y})$
4.  $D(X \pm Y) = D(X) + D(Y) \pm 2 \text{Cov}(X, Y)$

$$\frac{\text{(实际频数}-\text{理论频数})^2}{\text{理论频数}}$$

### 贝叶斯公式

因果关系  $A \rightarrow B$ , ,  $A_1, A_2, A_3, \dots, A_n$  是一个完备事件组。

$$P(B | A_i) = \frac{P(A_i B)}{P(B)} = \frac{P(A_i) P(B | A_i)}{\sum_{i=1}^n P(A_i) P(B | A_i)}$$

### 常用统计量

1. 样本均值:  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$

观察值:  $\bar{x} = \frac{1}{n} \sum_{i=1}^n X_i$

2. 样本方差

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

3. 样本标准差

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

观察值:  $X \rightarrow x$

4. 样本k阶 (原点) 矩

$$A_k = \frac{1}{n} \sum_{i=1}^n X_i^k, k = 1, 2, \dots$$

观察值:  $X \rightarrow x$

5. 样本k阶中心距

$$B_k = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^k, k = 2, 3, \dots$$



## 两个总体比率之差的检验

- 1. 假定条件
    - 两个总体都服从二项分布
    - 可以用正态分布来近似
  - 检验统计量
    - 检验  $H_0: \pi_1 - \pi_2 = 0$  
$$z = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$
    - 检验  $H_0: \pi_1 - \pi_2 = d_0$  
$$z = \frac{(p_1 - p_2) - d_0}{\sqrt{\frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2}}}$$
- $p = \frac{x_1 + x_2}{n_1 + n_2} = \frac{p_1 n_1 + p_2 n_2}{n_1 + n_2}$

ool of Software Engineering

## 3 总体比率和方差检验

### 两个总体比率之差的检验规则



假设	双侧检验	左侧检验	右侧检验
假设形式	$H_0: \pi_1 - \pi_2 = 0$ $H_1: \pi_1 - \pi_2 \neq 0$	$H_0: \pi_1 - \pi_2 \geq 0$ $H_1: \pi_1 - \pi_2 < 0$	$H_0: \pi_1 - \pi_2 \leq 0$ $H_1: \pi_1 - \pi_2 > 0$
统计量	$z = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$	$z = \frac{(p_1 - p_2) - d_0}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$	
拒绝域	$ z  > z_{\alpha/2}$	$z < -z_\alpha$	$z > z_\alpha$
P值决策	$P < \alpha$ 拒绝 $H_0$		