

# 李伯坚-统计学

公式、知识点整理

Rongxin Feng  
2021.7.17

第01章 绪论

目的：面对不确定的状况下，能够帮助我们做出明智决策的一种科学方法

- (1) 叙述统计
- (2) 推论统计：以样本推论总体

母体的特征：参数  
样本的特征：统计量

抽样：

坏的例子：兰登和罗斯福总统大选

抽样的方法：

随机抽样	简单随机抽样	每个总体的每个成员都编上号，来随机抽签
	系统抽样(systematic sampling)	每隔 k 个元素取一个样本
	分层抽样(stratified sampling)	依据总体中个体属性类分之互斥组别
	部落抽样(cluster sampling)	
非随机抽样	偶遇抽样(convenience sampling)	方便、费用低、时间省（误差相对大）市场调查
	配额抽样(quota sampling)	先分层，然后主观选择样本
	主观抽样(purposive sampling)	eg：对中部地区市场进行调查
	滚雪球抽样(snowball sampling)	请被访者介绍其朋友接受访问

效度 (validity)：测量工具（问卷等）可能影响测量结果的准确程度

内部效度：样本观察值师傅可信、可靠  
外部效度：研究结果能够普遍使用道样本来自的总体

尺度：

- 1 名义尺度：性别等
- 2.顺序尺度：连续型的
- 3.区间尺度：高中低
- 4.比率尺度：

常用问卷调查的区间尺度：

- 1. 常采用区间尺度
- 2. 李克特 (likert) 量表
- 3. 组距相等（非常满意、满意……非常不满意）

资料的集中趋势：

总体平均数(mu): $\mu$

样本平均数(average):  $\bar{x}$

中位数(median):

众数(mode):

钟形分布时:  $\bar{x}$ 、median、mode 合一

右偏: 从左往右依次是 mode、median、 $\bar{x}$ ;

左偏相反

资料的分散趋势:

全距(range): Max - Min

四分位居(Interquartile Range; IQR) :  $IQR = Q_3 - Q_1$ . (BOX-PLOT)

四分位差(Quartile Deviation; QD):  $QD = \frac{Q_3 - Q_1}{2}$

标准差(Standard Deviation):

$$\sigma^2 = \frac{\sum_{i=1}^n (x_i - \mu)^2}{N} \quad (\text{总体方差})$$

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} \quad (\text{样本方差})$$

变异系数(C.V) --coefficient of variance

方便比较两组资料的相对变异型:  $CV = \frac{\sigma}{\mu}$

## 第 02 章 柴比雪夫不等式

经验法则: 一个 $\sigma$ 68.3%, 两个 $\sigma$ 95.4%, 三个 $\sigma$ 99.7%

柴比雪夫不等式:

$$P(|X - \mu| \leq k\sigma) \geq 1 - \frac{1}{k^2} \quad (k > 1)$$

柴比雪夫单边版: 比马可夫不等式要精确一点

$$P(X \geq k) \leq \frac{\sigma^2}{\sigma^2 + k^2} \quad (\text{假设 } E(X) = 0, \text{VAR}(X) = \sigma^2)$$

马可夫不等式:

$$P(X \geq k) \leq \frac{E(X)}{k}$$

## 第 03 章 概率

排列: 从 n 个不同物品中选出 m 个排成一列

$$P_m^n = \frac{n!}{(n-m)!}$$

组合: 从 n 个不同物品中选出 m 个物品

$$C_n^m = \frac{n!}{m!(n-m)!} = \frac{n \cdot (n-1) \cdot \dots \cdot (n-m+1)}{m!}$$

二项式定理:

$$(x+y)^n = \sum_{k=0}^n C_k^n x^k y^{n-k}$$

$$\left(x + \frac{2}{x}\right)^8 = \sum_{r=0}^8 C_r^8 x^r \left(\frac{2}{x}\right)^{8-r} \quad (\text{当 } r \text{ 等于 } 6 \text{ 时就是 } x^6 \text{ 的系数})$$

概率:

样本空间 (sample space)

事件(Event)

$$P = \frac{\#(E)}{\#(\Omega)}$$

独立事件: 若  $P(A|B) = P(A)$  或  $P(B|A) = P(B)$ , 则称 A、B 两事件独立

互斥事件: 没有交集  $P(A \cap B) = 0$

条件概率:

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

例题: 一个家庭有 2 个小孩, 已知至少一个是男孩, 求两个均是男孩的概率

贝氏定理:

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

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

$$P(A_1|B) = \frac{P(B|A_1) \cdot P(A_1)}{\sum_i P(B|A_i) \cdot P(A_i)}$$

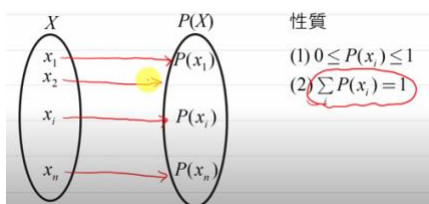
## 第 04 章 随机变数

### 随机变数 (random variable)

定义: 随机变数 X 是定义于样本空间的实数值函数

例子: 袋子中有 3 个红球, 2 个黑球, 设 X 表示红球的个数, X 的值的范围

概率函数:



随机变数的类型:

1. 离散型 (间断型) : 值域有限
2. 连续型: 值域无限

概率分布 (probability distribution) :

定义: 一个随机变数的可能值连同它的概率函数称为随机变数的概率分布。

离散型概率分布图:

Eg: 投掷两个骰子,  $X$  表示两个骰子的点数之和, 求概率分布图

$x$	2	3	4	5	6	7	8	9	10	11	12
$p(x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

连续型概率分布:

性质:

$$(1) P(X = x_i) = 0$$

$$(2) P(a < X < b) = \int_a^b f(x) dx$$

$$(3) P(-\infty < X < \infty) = 1$$

联合概率分布 (joint probability distribution) :

定义: 在一个随机试验中, 样本空间中有两个随机变数。可分别探讨概率分布, 也可探讨相互关系。

eg: 有 6 位顾客,  $X$  表示财力高中低分别用 1, 2, 3 表示;  $Y$  表示卖与不买分别用 1, 2 表示, 试求

联合概率分布

财力	购买
----	----

高	是
低	否
中	否
中	是
低	否
高	否

	1	2	
Y			
X			
1	1/6	1/6	2/6
2	1/6	1/6	2/6
3	0	2/6	2/6
	2/6	4/6	1

边际概率分布 (margin probability distribution)

$$P(X = 1,2,3) =$$

$$P(Y = 1,2,3) =$$

期望值 (expected value) :

定义: 经长期重复实验, 预期得到的平均是:

$$\text{间断型: } E(X) = \sum_{i=1}^n x_i P(X = x_i)$$

$$\text{连续型: } E(X) = \int_{-\infty}^{\infty} x \cdot f(x) dx$$

Eg: 袋子中有 10 元、5 元、1 元分别 1、3、4 枚, 求期望值

$x$	10	5	1
$P(X = x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{4}{8}$

$$E(X) = 10 \cdot (1/8) + 5 \cdot (3/8) + 1 \cdot (4/8)$$

期望值的线性性质:

$$E(aX) = \sum_{i=1}^n ax_i P(X = x_i) = a \sum_{i=1}^n x_i P(X = x_i) = aE(X)$$

$$E(aX + bY) = aE(X) + bE(Y)$$

$$E(a + bX) = E(a) + bE(X) = a + bE(X)$$

变异数 (variance) :

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

$$\text{Var}(X) = \sum_{i=1}^n (x_i - \mu)^2 P(X = x_i)$$

$$\text{Var}(X) = E[(x - \mu)^2]$$

定理:

$$\text{Var}(X) = E(x^2) - [E(x)]^2$$

定义: 间断型随机变异数, 离差平方和的平均

变异数的平方性质:

$$\text{Var}(aX) = a^2 \text{Var}(X)$$

$$\begin{aligned} \text{Var}(aX) &= E(a^2 X^2) - [E(aX)]^2 \\ &= a^2 E(X^2) - [aE(X)]^2 \\ &= a^2 E(X^2) - a^2 [E(X)]^2 = a^2 \text{Var}(X) \end{aligned}$$

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$

$$\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y)$$

## 第 05 章 间断型随机概率分布

### 负二项分布 (negative binomial distribution)

几何分布定义：做伯努利实验，成功概率为  $p$ ，**做到第一次成功停止**，所需的试验次数  $X$  的概率分布

$$\text{期望值和变异数: } E(x) = \frac{1}{p} \text{Var}(x) = \frac{q}{p^2}$$

负二项分布：做伯努利实验，成功概率为  $p$ ，**做到第  $k$  (正整数) 次成功停止**，所需的试验次数  $X$  的概率分布

设二项实验，成功概率为  $p$ ，失败概率为  $q$ ，设  $X$  表示第  $k$  次成功的次数，则概率分布为

(成, 成, 成, ..., 成, 成...败...败...败...败...败) 成

( $k-1$ )

( $x-k$ )

$$P(X = x) = C_{k-1}^{x-1} p^{k-1} q^{x-k} \cdot p = C_{k-1}^{x-1} p^k q^{x-k}$$

$$\text{期望值和变异数: } E(x) = \frac{k}{p} \text{Var}(x) = \frac{kq}{p^2}$$

### 超几何分布 (hyper geometric)

定义：设随机变量  $X$  为二项实验，但每次**抽样后不放回**，则称  $X$  为超几何分布

Eg: 设袋子中共  $N$  个球，红球  $k$  个，抽后不放回，现在随机抽出  $n$  个球，求红球个数  $x$  的概率

红球  $k$  个

非红球:  $N-k$  个

$$P(X = x) = \frac{C_x^k C_{n-x}^{N-k}}{C_n^N}$$

$$\text{期望值: } E(X) = \frac{n(n-1)k(k-1)}{N(N-1)}$$

$$\text{变异数: } \text{VAR}(X) = \frac{(N-n)}{(N-1)} \cdot n \frac{k}{N} \left(1 - \frac{k}{N}\right)$$

### 二项分布

二项实验/伯努利实验的定义：在  $n$  次实验中，实验结果只有成功和失败两种可能，每次实验成功或失败的概率都相同（总体= $N$ ，每次抽样均放回），每次实验均相互独立，则称此事件为二项实验或伯努利实验。

$$P(X = x) = C_x^n p^x q^{n-x}$$

二项分布的期望值  $E(X) = np$

二项分布的变异数  $\text{VAR}(X) = npq$

例题：一个箱子中有 10 个灯泡，已知有 2 个不良品，以抽后放回的方式抽 3 次，令 X 表示抽出不良品个数，试求概率分配

$x$	0	1	2	3
$p(x)$	$C_0^3 \left(\frac{2}{10}\right)^0 \left(\frac{8}{10}\right)^3$	$C_1^3 \left(\frac{2}{10}\right)^1 \left(\frac{8}{10}\right)^2$	$C_2^3 \left(\frac{2}{10}\right)^2 \left(\frac{8}{10}\right)^1$	$C_3^3 \left(\frac{2}{10}\right)^3 \left(\frac{8}{10}\right)^0$

#### 超几何分布与二项分布的比较

- (1) 取后放回  $\rightarrow$  二项分布
- (2) 取后不放回  $\rightarrow$  超几何分布
- (3) 二项分布的期望值 = np, 超几何分布的期望值=np
- (4) 二项分布的变异数 = npq, 超几何分布的变异数= $\frac{N-n}{N-1}npq$ 。当  $n \ll N$  时，超几何分布接近二项分布。

#### 第 06 章泊松分布 (poisson distribution)

探讨一段时间内事件的发生次数  $\lambda$

性质：

1. 一段时间发生的次数与另外一段时间发生的次数独立
2. 一段时间发生的平均次数与时间长短成比例
3. 在极短的时间内发生的概率趋近于 0

泊松分布的期望值与变异数

泊松分布可以看作当  $n \rightarrow \infty$ ,  $p \rightarrow 0$  时的二项分布

$$E(X) = np = \lambda$$

$$VAR(X) = np = \lambda$$

#### 第 07 章统计相关理论工具

##### 动差的概念

K 阶动差 (moment) 的观念

$$\text{中心动差: } \mu_k = E[(X - \mu)^k] = \begin{cases} \sum_x (x - \mu)^k \cdot P(x) \\ \int_{-\infty}^{\infty} (x - \mu)^k \cdot f(x) dx \end{cases}$$

$$\text{原点动差: } \mu'_k = E[(X - 0)^k] = \begin{cases} \sum_x x^k \cdot P(x) \\ \int_{-\infty}^{\infty} x^k \cdot f(x) dx \end{cases}$$



一阶动差:  $\mu'_1 = E[X] = \begin{cases} \sum_x x \cdot P(x) \\ \int_{-\infty}^{\infty} x \cdot f(x) dx \end{cases} = \mu$  找寻随机变量的[中心], 也就是期望值

二阶动差  $\mu_2 = E[(X - \mu)^2] = \begin{cases} \sum_x (x - \mu)^2 \cdot P(x) \\ \int_{-\infty}^{\infty} (x - \mu)^2 \cdot f(x) dx \end{cases}$  测量随机变量与[中心]的距离平方, 变异数的定义

三阶动差:  $\mu_3 = E[(X - \mu)^3] = \begin{cases} \sum_x (x - \mu)^3 \cdot P(x) \\ \int_{-\infty}^{\infty} (x - \mu)^3 \cdot f(x) dx \end{cases}$  有正有负, 测量随机变量与[中心]的距离立方, 可测偏态

偏态 (skewness)

$$\text{偏态系数} = \frac{\mu_3}{\sigma^3}$$

右偏: 有长长的右尾, 资料右端有较多的极值

左偏: 有长长的左尾, 资料左端有较多的极值

四阶动差:  $\mu_4 = E[(X - \mu)^4] = \begin{cases} \sum_x (x - \mu)^4 \cdot P(x) \\ \int_{-\infty}^{\infty} (x - \mu)^4 \cdot f(x) dx \end{cases}$  必定为正值, 测量随机变量与[中心]的距离 4 次方, 可测峰度

峰度 (kurtosis)

$$\text{峰度系数} = \frac{\mu_4}{\sigma^4} \quad (>3 \text{ 为高峽峰}, \quad = 3 \text{ 为常态峰}, <3 \text{ 为低阔峰})$$

动差生成函数 (moment- generating function)

$$M_X(t) = E(e^{tx})$$

$$(1) M_X(t) = \sum_{i=1}^n e^{tx_i} P(X = x_i) \quad \text{间断型}$$

$$(2) M_X(t) = \int_{-\infty}^{\infty} e^{tx} \cdot f(x) dx \quad \text{连续型}$$

动差生成函数在以上级数或积分收敛时才会存在

$$\begin{aligned} \frac{d^r M_X}{dt^r} &= \sum_{i=1}^n x_i^r e^{tx_i} P(X = x_i) \\ \left. \frac{d^r M_X}{dt^r} \right|_{t_i} &= \sum_{i=1}^n x_i^r P(X = x_i) = \mu'_r \quad r \text{ 阶原点动差} \end{aligned}$$

$$\begin{aligned} \frac{d^r M_X}{dt^r} &= \int_{-\infty}^{\infty} e^{tx} x^r \cdot f(x) dx \\ \left. \frac{d^r M_X}{dt^r} \right|_{t=0} &= \int_{-\infty}^{\infty} x^r \cdot f(x) dx = \mu'_r \quad r \text{ 阶原点动差} \end{aligned}$$

## 第 08 章连续型随机变数 (用 histogram)

$$\text{期望值: } E(X) = \int_{-\infty}^{\infty} x \cdot f(x) dx$$

$$\text{变异数: } \text{Var}(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx = \int_{-\infty}^{\infty} x^2 f(x) dx - \mu^2$$

### 伽玛函数

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} dx \quad \alpha > 0$$

### 伽玛函数性质

$$\text{递回式: } \Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$$

$$\Gamma(n) = (n - 1)(n - 2) \cdots \Gamma(1)$$

$$\Gamma(1) = 1$$

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$$

$$\begin{aligned} \Gamma\left(\frac{1}{2}\right) &= \int_0^{\infty} x^{-\frac{1}{2}} e^{-x} dx, \quad \text{令 } x=u^2, \quad dx = 2u du \\ &= \int_0^{\infty} u^{-1} e^{-u^2} 2u du \\ &= 2 \int_0^{\infty} e^{-u^2} du \\ &= 2 \frac{\sqrt{\pi}}{2} \end{aligned}$$

### 伽玛分布

$$f(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\beta}}, \alpha > 0, \beta > 0$$

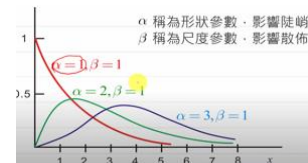
(1)  $\alpha$  是形状参数, 影响陡峭程度,

$\alpha=1$  为递减函数,  $\alpha$  越大越呈钟形

(2)  $\beta$  是尺度参数, 影响散布程度

$$\text{期望值: } \mu = \alpha\beta$$

$$\text{变异数: } \text{VAR}(x) = \alpha\beta^2$$



### 指数分布:

指数分布是伽玛分布  $\alpha = 1$  的特例, 概率

密度函数为:

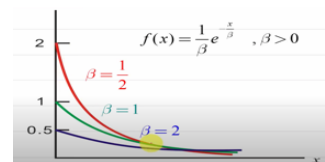
$$f(x) = \frac{1}{\beta} e^{-\frac{x}{\beta}}, \beta > 0, \text{ 又记作 } f(x) =$$

$$\lambda e^{-\lambda x}, \lambda > 0$$

$$\left(\frac{1}{\beta} = \lambda\right)$$

$$\text{期望值: } E(X) = \beta$$

$$\text{变异数: } \text{VAR}(x) = \beta^2$$



概率的算法：

$$P(X > a) = \int_a^{\infty} \frac{1}{\beta} e^{-\frac{x}{\beta}} dx = e^{-\frac{a}{\beta}}$$

指数分布的用途：

1. 排队理论：电话间隔、两个顾客结账间隔时间
2. 零件寿命

排队理论例子：

假设医院门诊，看病的平均时间为 $\mu$ ，一位病人看病结束后，门口亮灯请下一位病人进来，则亮灯的平均时间是 $\mu$ 。若将灯亮视为一个事件发生，则亮灯过程近似于柏松过程。假设 T 表示两次亮灯时间的间隔，则 T 服从指数分布。

指数分布的性质：

无记忆性质（几何分布也具有无记忆性质）

$$P(T > a + b | T > a) = P(T > b), a > 0, b > 0$$

卡方分布：

$$\text{伽玛分布: } f(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\beta}}, \alpha > 0, \beta > 0$$

当 $\alpha = \frac{\nu}{2}$ ,  $\beta = 2$  时其概率密度函数为：卡方分

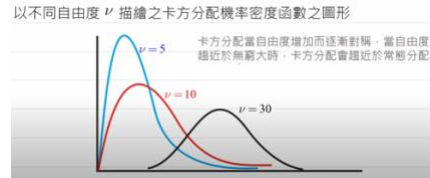
布

$$f(x) = \frac{1}{2^{\frac{\nu}{2}} \Gamma(\frac{\nu}{2})} x^{\frac{\nu}{2}-1} e^{-\frac{x}{2}}, x > 0$$

其中 $\nu$ 为正整数，称作自由度，称为以 $\nu$ 为自由度(degree of freedom)之卡方分布(Chi-Square distribution)

期望值： $E(X) = \nu$

变异数： $\text{VAR}(X) = 2\nu$



卡方分布的性质：

两个卡方分布之和

设  $X_1$  与  $X_2$  为两组相互独立的卡方分布，其自由度分别为 $\nu_1$ 和 $\nu_2$ ，则  $Y = X_1 + X_2$  服从自由度为 $\nu_1 + \nu_2$ 的卡方分布

定理： $Z^2 \sim \chi_1^2$

常用卡方统计量

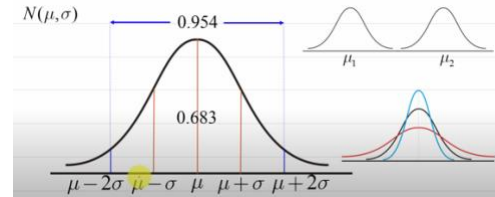
$$\frac{(n-1)s^2}{\sigma^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} \sim \chi_{n-1}^2$$

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

## 第 09 章 正态分布

### 正态分布

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad \text{图形以}\mu\text{为中心之钟形}$$



### 标准化

$$Z = \frac{X-\mu}{\sigma} \quad Z \sim N(0,1)$$

期望值:  $E(X) = \mu$

变异数:  $\text{VAR}(X) = \sigma^2$

标准正态分布:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \quad \text{其概率 } P(a < X < b) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$

### 二项分布与正态分布

二项分布中, 若实验次数  $n$  越大时, 则无论成功概率  $p$  如何, 其分布越来越接近正态分布

当二项分布  $np > 5$  &  $nq > 5$ , 可以以正态分布来估计概率。

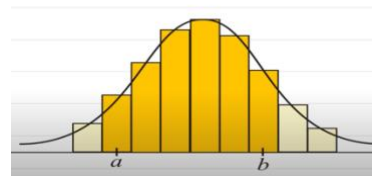
连续校正因子(correction for continuity)把离散型的

改为连续型来计算

$$P(a \leq X \leq b) = P(a - 0.5 < x < b + 0.5)$$

$$P(X \geq A) = P(x > a - 0.5)$$

$$P(a \leq X) = P(x < a + 0.5)$$



## 第 10 章 联合概率分布

### 联合概率分布

假设  $X$  和  $Y$  为间断型随机变数, 以  $P(X = a, Y = b)$  假设  $X = a$  且  $Y = b$  发生的概率, 所有可能发生的概率形成的联合概率质量函数 (joint probability mass function) 为  $f(x, y)$

$$f(a, b) = P(x = a, Y = b)$$

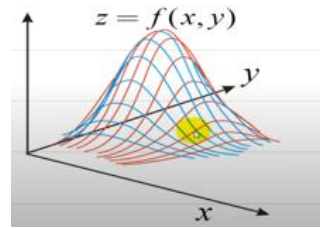
$$0 \leq f(a, b) \leq 1$$

$$\sum_x \sum_y f(x, y) = 1$$

联合概率密度函数 (连续型 joint probability density function)

$$0 \leq f(a, b) \leq 1$$

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1$$



### 边缘概率分布函数

假设假设  $X$  和  $Y$  为间断型随机变数, 若  $P(X = a, Y = b)$  为联合概率分布, 则

$$\sum_y P(x, y) \text{ 表示 } X \text{ 的边缘分布函数}$$

$$\sum_x P(x, y) \text{ 表示 } Y \text{ 的边缘分布函数}$$

假设假设  $X$  和  $Y$  为连续型随机变数, 若  $f(X = a, Y = b)$  为联合概率分布, 则

$$f_x(x, y) \text{ 表示 } X \text{ 的边缘概率密度函数}$$

$$f_y(x, y) \text{ 表示 } Y \text{ 的边缘概率密度函数}$$

若  $f(x, y) = f_x(x) * f_y(y)$ , 则  $X$  和  $Y$  独立, 否则相依

若  $X$  和  $Y$  随机变数相互独立, 则其联合概率可由两个边缘概率相乘

若  $X$  和  $Y$  随机变数相依, 则只能用联合概率描述其共同行为

### 联合概率分布的期望值

设  $X$  和  $Y$  为随机变数  $f(x, y)$  为联合概率分布函数, 设  $g(x, y)$  为  $R^2 \rightarrow R$  实值函数, 函数  $g(x, y)$  的期望值为  $E[g(x, y)]$

$$E[g(x, y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f(x, y) dx dy$$

$$E[g(x, y)] = \sum_x \sum_y g(x, y) f(x, y)$$

$$E(x) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f(x, y) dy dx = \int_{-\infty}^{\infty} x f_x(x) dx$$

$$\text{VAR}(x) = E(X^2) - [E(X)]^2$$

### 协方差 (共变数 covariance)

设  $X$  和  $Y$  为随机变数,  $E(X) = \mu_1$ ,  $E(Y) = \mu_2$ ,  $X$  和  $Y$  的协方差为:

$$\text{Cov}(X, Y) = E[(X - \mu^1)(Y - \mu^2)] = E(XY) - E(X) E(Y)$$

### 协方差的正负

协方差的值  $(-\infty, \infty)$ , 无法作为线性关系的绝对度量

$X$  和  $Y$  的协方差  $> 0$ , 表示  $X$  和  $Y$  同方向变动

X 和 Y 的协方差 $<0$ , 表示 X 和 Y 反方向变动

X 和 Y 的协方差 $=0$ , 表示 X 和 Y 没有线性关系, 但不代表没有其他关系

协方差与独立之间的关系

若 X 和 Y 相互独立 $\Rightarrow \text{Cov}(X, Y) = 0$ , 反过来不成立

条件变异数

双重期望值定理:  $E[E(x|y)] = E(X)$  说明了随机变数  $E(X|Y)$  和 X 有相同的期望值

条件变异数定理  $\text{Var}(X) = E[\text{Var}(x|y)] + \text{Var}[E(x|y)]$

条件变异数不等式:  $\text{Var}(X) \geq \text{Var}[E(x|y)]$

## 第 11 章 抽样分布一

抽样分布的观念

统计量

推论统计: 有未知总体抽出样本, 再由抽出的样本找出特征推断总体的特性, 样本所显示的特征叫作统计量

1. 统计量为随机变数, 故也有其分布, 称为抽样分布

eg:  $\bar{X}$  的抽样分布称为平均数的抽样分布;

$\hat{p}$  的抽样分布称为样本比例的抽样分布

$S^2$  的抽样分布称为样本方差的抽样分布

2. 抽样分布的标准差, 习惯称为标准误差

样本平均数抽样分布

由一大小为 N 的总体抽出一组样本  $x_1 + x_2 + \dots + x_n$  则样本平均数  $\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$  则

$$E(\bar{x}) = \mu \quad \mu_{\bar{x}} = \mu$$

$$\text{Var}(\bar{x}) = \frac{\sigma^2}{n} \quad \sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$$

$\bar{X}$  的抽样分布

常态总体	大样本	$\bar{x} \sim N(\mu, \frac{\sigma^2}{n})$
	小样本 ( $\sigma$ 已知)	$\bar{x} \sim N(\mu, \frac{\sigma^2}{n})$
	小样本 ( $\sigma$ 未知)	$\bar{x} \sim t(\mu, \frac{s^2}{n})$
非常态总体	大样本 C.L.T	$\bar{x} \sim N(\mu, \frac{\sigma^2}{n})$
	大样本 Nonparametric	视总体状况

样本比例抽样分布

与二项分布的观念一致，即成功与失败两种状况

符号：总体比例  $p$  样本比例  $\hat{p}$

$$\text{求法: } P = \frac{x}{N} \quad \hat{p} = \frac{x}{n}$$

$$E(\hat{p}) = P$$

$$\text{Var}(\hat{p}) = \frac{pq}{n}$$

大样本( $np > 5$ & $nq > 5$ )	$\hat{p} \sim N(p, \frac{pq}{n})$
小样本（无限总体）放回	$\hat{p} \sim \text{binomial}(p, \frac{pq}{n})$
小样本（有限总体）不放回	$\hat{p} \sim \text{Hypergeometry}(p, \frac{pq}{n} * \frac{N-n}{N-1})$

样本变异数的抽样分布

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

$$\frac{(n-1)s^2}{\sigma^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} \sim \chi^2_{(n-1)} \quad (\text{自由度为 } n-1 \text{ 的卡方分布})$$

#### 中央极限定理(C.L.T)

假设  $x_1, x_2 \dots x_n$  是独立且同分布的随机变数，期望值  $\mu$ ，变异数  $\sigma^2$ ，则当样本数量增大时，其样本平均数趋于正态分布，以数学符号表示：

$$\lim_{n \rightarrow \infty} \frac{x_1 + x_2 + \dots + x_n}{n} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad n \text{ 超过 } 30 \text{ 即可}$$

#### 标准差和标准误差

标准差（standard deviation）：是用来描述资料的分散程度

标准误（standard error of the mean）：是我们凭借手边的样本（sample）资料，对总体（population）的平均值做估计时，对这个估计结果误差程度的表示方法

$$\text{总体平均数 } \mu = \frac{x_1 + x_2 + \dots + x_N}{N} \quad \text{总体标准差 } \sigma = \sqrt{\frac{\sum_{i=1}^N (x_i - \mu)^2}{N}}$$

$$\text{样本平均数 } \bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} \quad \text{样本标准差 } s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$$

$$\text{标准误 } SE = \frac{s}{\sqrt{n}}$$

## 第 12 章 抽样分布二

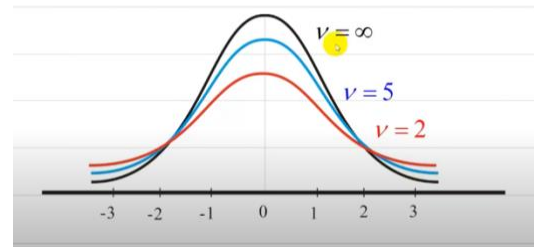
### T 分布

当  $n \geq 30$  时, 若以  $s^2$  取代  $\sigma^2$ , 统计量  $\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$  以  $\frac{\bar{x}-\mu}{s/\sqrt{n}}$  取代, 接近标准正态分布 (中央极限定理), 但当  $n < 30$  时, 我们以 T 分布表示:

$$T = \frac{\bar{x}-\mu}{s/\sqrt{n}}$$
$$T = \frac{\bar{x}-\mu}{s/\sqrt{n}} = \frac{Z}{\chi^2/(n-1)}, \text{ 其中 } \chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

T 分布的自由度为  $\nu$ , 其概率密度函数为:

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



求 T 分布概率密度函数步骤

已知: T 分布是标准正态分布与卡方分布的组合

1. 写出两分布的联合概率函数
2. 利用变数变换, 加入 jacobian
3. 利用边际概率函数求出 T 分布概率密度函数

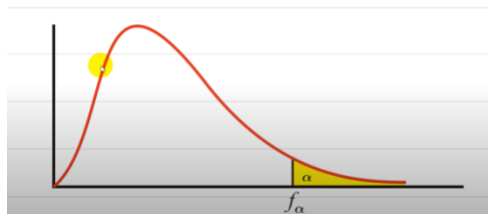
### F 分布

定义: 两个卡方分布各自除以自由度之比值称为 F 分布

$$F = \frac{\chi_1^2/\nu_1}{\chi_2^2/\nu_2}$$

其概率密度函数:

$$h(f) = \frac{\Gamma(\frac{\nu_1+\nu_2}{2})\left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}}}{\Gamma(\frac{\nu_1}{2})\Gamma(\frac{\nu_2}{2})} \frac{f^{\frac{\nu_1}{2}-1}}{\left(1+\frac{\nu_1}{\nu_2}f\right)^{\frac{(\nu_1+\nu_2)}{2}}}$$





右偏 + 有 2 个自由度

F 分布的性质

1. F 分布  $f(v_1, v_2)$  与  $\frac{1}{f(v_2, v_1)}$  分布相同
2.  $f_{1-\alpha}(v_2, v_1) = \frac{1}{f_\alpha(v_1, v_2)}$

F 分布的期望值

$$E(F) = E\left(\frac{\chi_1^2/v_1}{\chi_2^2/v_2}\right) = \frac{v_2}{v_2 - 2}$$

各种分布之间的关系

1. Z 的定义  $Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$
2. t 的定义  $t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{Z}{\sqrt{\chi^2/(n-1)}}$
3. F 的定义 :  $F = \frac{\chi_1^2/(n_1-1)}{\chi_2^2/(n_2-1)} = F(n_1 - 1, n_2 - 1)$
4. 定理:  $Z^2 \sim \chi^2(1)$

## 第 13 章 假设检验导论

假设检验导论

1. 对总体参数设定一假设
2. 利用由样本所获得的样本统计量, 以检验总体参数是否符合假设
3. 对此假设做出决策, 也就是接受或拒绝此假设

假设:

1. 依据检验内容建立虚无假设(Null Hypothesis):  $H_0$
2. 依据检验被容建立对立假设(Alternative Hypothesis):  $H_1$

误差:

型 1 误差 (Type 1 error) : 以  $\alpha$  表示  $\alpha = P(\text{拒绝}|\text{为真})$  --- 显著水平  
型 2 误差 (Type 2 error) : 以  $\beta$  表示  $\beta = P(\text{接受}|\text{不真})$  ---- 检定力  
检定力 =  $1 - \beta$

步骤:

1. 依据检验内容建立虚无假设  $H_0$
2. 依据检验被容建立对立假设  $H_1$  (根据拒绝域分为单尾和双尾)
3. 确定显著水平  $\alpha$
4. 判断检验方法 (z 分布、t 分布、卡方分布、f 分布)
5. 建立拒绝域 (接受域)  $\rightarrow$  得出结论

控制误差：

为了降低型 1 误差 $\alpha$ 有以下 2 种方法

1. 加宽接受域，会导致 $\beta$ 增加，牺牲了型 2 误差
2. 加大样本数， $\alpha$ 和 $\beta$ 都会降低，增加工资成本

## 第 14 章 点估计一

### 估计导论

估计的目的

1. 推论统计学：利用样本统计量及其抽样分布来估计总体的特性
2. 参数估计(parametric statistics)：假设总体为常态或样本为大样本的情况
3. 非参数估计(nonparametric statistics)：总体分布不明或样本为小样本
4. 分为统计估计(statistics estimation) 和 假设检验(Hypothesis testing)

估计的方法：

1. 点估计(point estimation)：样本平均数估计总体期望值
2. 区间估计(interval estimation)：利用样本统计量估计出一个区间去估计总体参数，此区间可靠度可知。eg：台北市每月大学生零花钱在 95%的置信水平下，介于[4000,7000]

估计的评价标准：

1. 不偏性(Unbiasedness)
2. 有效性(Efficiency)
3. 一致性(Consistency)
4. 充分性(Sufficiency)

总体平均数 $\mu$ 的估计：样本中位数、样本平均数（常用）

总体变异数 $\sigma^2$ ，以样本变异数  $S^2$ 最为常用，因其具有不偏性

### 常用的估计方法

1. 矩估计/动差法(method of moments estimation)
2. 最大似然值(Maximum Likelihood Estimation, 简作 MLE)
3. 贝叶斯估计法(Bayesian estimation)

1. 矩估计/动差法(method of moments estimation)

大数法则(Law of large number)

重复实验所获得的一组随机样本 $x_1, x_2 \cdots x_n$ ，均为独立且同分布，当试验次数  $n \rightarrow \infty$  时，则样本平均是会趋近于期望值。

$$\mu'_k = E(x_i^k) = \begin{cases} \sum_x x^k \cdot P(x) \\ \int_{-\infty}^{\infty} x^k \cdot f(x) dx \end{cases} : k \text{ 階動差} \quad m'_k = \frac{1}{n} \sum_{i=1}^n x_i^k : k \text{ 階樣本動差}$$

K 阶动差是 X 的  $k^2$  的期望值，使用 X 的  $k^2$  的样本平均数来估计称为动差法

$$\text{一阶动差: } \mu'_1 = E[X] = \begin{cases} \sum_x x \cdot P(x) \\ \int_{-\infty}^{\infty} x \cdot f(x) dx \end{cases} = \bar{x}$$

$$\text{二阶动差 } \mu_2 = E[(X - \mu)^2] = \begin{cases} \sum_x (x - \mu)^2 \cdot P(x) \\ \int_{-\infty}^{\infty} (x - \mu)^2 \cdot f(x) dx \end{cases}$$

## 2. 最大似然值(Maximum Likelihood Estimation, 简作 MLE)

原理：从总体中抽出一组随机样本，总体的参数  $\theta$  是我们想要研究的对象，若能从样本中找到一个估计值  $\hat{\theta}$ ，使得这组样本发生的可能性最大。则此估计值  $\hat{\theta}$  为  $\theta$  的最大似然估计值。

概似函数：

假设  $x_1, x_2 \dots x_n$  为抽自总体  $f(X; \theta)$  的一组随机样本，则其概似函数 (Likelihood Estimation) 定义为  $L(\theta)$ ；这就是  $n$  个随机变数  $x_1, x_2 \dots x_n$  的联合概率分布  $L(\theta) = f(X_1, X_2, \dots X_n; \theta)$

$$L(\theta) = f(X_1, X_2, \dots X_n; \theta) = f(X_1; \theta) f(X_2; \theta) \dots f(X_n; \theta) = \prod_{i=1}^n f(X_i; \theta)$$

步骤：

- 建立  $L(\theta) = f(X_1, X_2, \dots X_n; \theta)$
- 取对数  $\ln L(\theta) = \ln f(X_1, X_2, \dots X_n; \theta)$  变成单调增函数
- 求一阶导数并令为 0  $\frac{d \ln(L(\theta))}{d\theta} = 0$
- 以二阶导数判别是否为极大  $\frac{d^2 \ln(L(\theta))}{d\theta^2} < 0$

## 第 15 章 点估计二

### 3. 贝叶斯估计法(Bayesian estimation)

$$\text{条件概率: } P(B | A) = \frac{P(A \cap B)}{P(A)}$$

$$\text{贝叶斯定理: } P(A | B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A) \cdot P(B|A)}{P(A)P(B|A) + P(A')P(B|A')}$$

贝叶斯估计法：

1. 若我们想要对  $\theta$  进行估计，首先假设有一事前概率，此概率视为主观概率
2. 取样后，根据观测结果，利用贝叶斯定理修正对  $\theta$  的估计，如此便得到事后概率是利用贝叶斯定理而得
3. 事后概率是一条件概率
4. 事后概率与事前概率可能会有很大的差异

不偏估计量 (unbiasedness)

定义：若  $E(\hat{\theta}) = \theta$ ，则称  $\hat{\theta}$  为  $\theta$  的不偏估计量

$\bar{x}$  是  $\mu$  的不偏估计量

$S^2$  是  $\sigma^2$  的不偏估计量

有效估计量 (efficiency)

定义: 若  $E(\hat{\theta}_1) = \theta$ ,  $E(\hat{\theta}_2) = \theta$ , 则称  $\hat{\theta}_1$  与  $\hat{\theta}_2$  估计量都是  $\theta$  的不偏估计量; 若  $Var(\hat{\theta}_1) < Var(\hat{\theta}_2)$ , 则  $\hat{\theta}_1$  比  $\hat{\theta}_2$  有效

一致估计量 (consistency)

定义: 当样本数增大时, 估计值与总体的差异会越来越小, 当样本数趋近于  $\infty$  时, 差异趋近于 0

$$\lim_{n \rightarrow \infty} Var(\hat{\theta}_n) = 0$$

充分估计量 (sufficiency)

定义: 若点估计  $\hat{\theta}$  能充分表达出样本资料中有关总体参数  $\theta$  的资讯,  $\hat{\theta}$  称为  $\theta$  的充分估计量

## 第 16 章 单个总体平均数假设检验

## 第 17 章 单个总体比例假设检验

与二项分布的观念一致, 即成功与失败两种状况

符号: 总体比例  $p$  样本比例  $\hat{p}$

$$\text{求法: } P = \frac{X}{N} \quad \hat{p} = \frac{x}{n}$$

$$E(\hat{p}) = P$$

$$Var(\hat{p}) = \frac{pq}{n}$$

总体变异数假设检验

总体变异数:  $\sigma^2$  样本变异数:  $s^2$

只要一样本数为  $n$  的简单随机样本是选自一正态总体, 则  $\chi^2 = \frac{(n-1)s^2}{\sigma^2}$  的抽样分布服从自由度为  $n-1$

的卡方分布

总体比例假设检验

符号: 总体比例  $p$  样本比例  $\hat{p}$ , 大样本是用  $Z$  分布来做  $z = \frac{\bar{x} - \mu}{s/\sqrt{n}} \quad z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$

## 第 18 章 双总体假设检验一

### 两个总体平均数之差假设检验(独立大样本)

设  $X_1$  和  $X_2$  分别表示两个总体, 若  $X_1$  和  $X_2$  统计独立, 否则为非独立总体

$$E(\bar{x}_1 - \bar{x}_2) = \mu_1 - \mu_2$$

$$Var(\bar{x}_1 - \bar{x}_2) = Var(\bar{x}_1) + Var(\bar{x}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

分三种情况:

1. 两总体变异数已知  $\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

2. 两总体变异数未知  $\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

3. 两总体变异数未知, 但确认相等  $\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$  (加权的做法)

加权样本标准差

$$\text{第一组样本: } s_1^2 = \frac{\sum_{i=1}^{n_1} (x_{1i} - \bar{x}_1)^2}{n_1 - 1}$$

$$\text{第二组样本: } s_2^2 = \frac{\sum_{i=1}^{n_2} (x_{2i} - \bar{x}_2)^2}{n_2 - 1}$$

$$\text{混合样本标准差 } s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

两个总体平均数之差假设检验(独立小样本)

独立小样本 ( $n_1 < 30, n_2 < 30$ )

1. 总体变异数  $\sigma_1^2$  和  $\sigma_2^2$  已知  $z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{x}_1 - \bar{x}_2}} \quad \sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
2. 总体变异数  $\sigma_1^2$  和  $\sigma_2^2$  未知, 确认不相等  $t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_{\bar{x}_1 - \bar{x}_2}} \quad s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$
3. 两总体变异数未知, 但确认相等  $t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_{\bar{x}_1 - \bar{x}_2}} \quad \sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$  其中  $s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$

两个总体平均数之差假设检验(相依样本)

- (1) 令  $D_i = X_{1i} - X_{2i}$  视为同一样本处理
- (2) 一样有大样本、小样本、单尾、双尾检验

- |   |   |
|---|---|
| (A) 大样本总体变异数已知 $Z = \frac{\bar{D} - \mu_D}{\sigma_D}$ | (C) 小样本总体常态变异数已知 $Z = \frac{\bar{D} - \mu_D}{\sigma_D}$ |
| (B) 大样本总体变异数未知 $Z = \frac{\bar{D} - \mu_D}{s_D}$      | (D) 小样本总体常态变异数未知 $t = \frac{\bar{D} - \mu_D}{s_D}$      |

## 第 19 章 双总体假设检验二

基本假设

- (1) 第一个总体之比例  $P_1$  第二个总体之比例  $P_2$
- (2) 第一个总体抽出之样本之比例  $P_1$  · 第二个总体抽出之样本之比例  $P_2$
- (3) 总体之比例差  $p_1 - p_2$  · 样本之比例差  $\hat{p}_1 - \hat{p}_2$
- (4) 大部分以大样本为主 · 根据中央极限定理 (CLT) · 抽样分配

$$\begin{aligned} \hat{p}_1 &\sim N\left(p_1, \frac{p_1 q_1}{n_1}\right) & \hat{p}_2 &\sim N\left(p_2, \frac{p_2 q_2}{n_2}\right) \\ \hat{p}_1 - \hat{p}_2 &\sim N\left(P_1 - P_2, \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}\right) \end{aligned}$$

总体比例之差检验统计量

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{s_{\hat{p}_1 - \hat{p}_2}}$$

其中  $s_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$  因为  $p_1$  和  $p_2$  未知, 故以  $\hat{p}_1$  和  $\hat{p}_2$  估计

若虚无假设  $H_0: P_1 = P_2$  时,  $s_{\hat{p}_1 - \hat{p}_2} = \sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$  这里的  $p$  以加权的方式估计  $p = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}$

### 两个总体变异数之比例假设检验

- (1) 第一个总体变异数 $\sigma_1^2$ , 第二个总体变异数 $\sigma_2^2$
- (2) 第一个样本变异数 $s_1^2$ , 第二个样本变异数 $s_2^2$
- (3) 统计量 $\frac{\frac{s_1^2}{\sigma_1^2}}{\frac{s_2^2}{\sigma_2^2}} = \frac{\frac{s_1^2}{\sigma_1^2}}{\frac{s_2^2}{\sigma_2^2}} \sim F(n_1-1, n_2-1)$

### McNemar 檢定--相依样本比例差检验

McNemar 检验又称非独立样本比率数的卡方检验, 或称为相依样本的卡方检验

- McNemar 检验采用 Z 分布检验是在大样本的前提下
- McNemar 检验针对相依样本比例差检验
- B+C 最低要求>10

主要探讨事件发生前后「由是与否」与「由否为是」的个数是否相等, 也就是检验事件发生前后比例是否更改

检验的方法:

- (1) 以 2 阶列联表最为清楚
- (2)

事件后 事件前	成功	失败
	成功	失败
成功	A ( $P_1$ )	B( $P_2$ )
失败	C( $P_3$ )	D( $P_4$ )

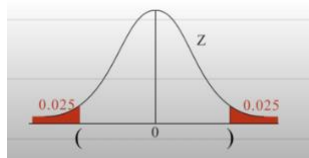
- (3) 虚无假设  $H_0: P_2 = P_3$ ;  $H_1: P_2 \neq P_3$  |  $H_1: P_2 > P_3$  |  $H_1: P_2 < P_3$
- (4) 检验统计量 $z = \frac{B-C}{\sqrt{B+C}}$

### 第 20 章 单个总体区间估计

#### 总体平均数之置信区间 (大样本)

区间估计概念

1. 区间估计是以点估计量的抽样分布为基础
2. 从区间范围来估计总体平均数 $\mu$ 的落点区域
3.  $\mu$ 的 $(1-\alpha)\%$ 的置信区间的意义为: 我们有 95%的信心/把握/概率 $\mu$ 会落在此区间



$\bar{x}$ 之抽样分布与置信区间

(1) 当总体变异数 $\sigma^2$ 已知时, 统计量  $z_{\alpha/2} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$

$$P(-Z_{\alpha/2} < \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} < Z_{\alpha/2}) = 1 - \alpha$$

$$P(\bar{X} - Z_{\alpha/2} \cdot \sigma/\sqrt{n} < \mu < \bar{X} + Z_{\alpha/2} \cdot \sigma/\sqrt{n}) = 1 - \alpha$$

(2) 当总体变异数 $\sigma^2$ 未知时, 以 s 取代即可 (原本应以 t 值取代, 因为大样本可以用 Z)

#### 总体平均数之置信区间 (小样本)

(1) 当总体变异数 $\sigma^2$ 已知时, 统计量  $z_{\alpha/2} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$

$$P(-Z_{\alpha/2} < \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} < Z_{\alpha/2}) = 1 - \alpha$$

$$P(\bar{X} - Z_{\alpha/2} \cdot \sigma/\sqrt{n} < \mu < \bar{X} + Z_{\alpha/2} \cdot \sigma/\sqrt{n}) = 1 - \alpha$$

(2) 当总体变异数 $\sigma^2$ 未知时, 以 s 取代即可 (以 t 修正)

$$P\left(\bar{X} - Z_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} < \mu < \bar{X} + Z_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}\right) = 1 - \alpha \quad (\bar{X} - t_{\alpha/2} \cdot s/\sqrt{n}, \bar{X} + t_{\alpha/2} \cdot s/\sqrt{n})$$

#### 总体比例之置信区间

大样本以 Z 分布处理, 小样本以二项分布处理

(1) 当总体比例 P 的点估计 $\hat{p}$ 已经求出后

(2) 根据中央极限定理,  $\hat{p}$ 的抽样分布 $\hat{p} \sim N(P, \frac{pq}{n})$

$$(3) \quad z_{\alpha/2} = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$

$$(4) \quad \text{置信区间}(\hat{p} - z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p} \cdot \hat{q}}{n}}, \hat{p} + z_{\alpha/2} \cdot \sqrt{\frac{\hat{p} \cdot \hat{q}}{n}})$$

#### 总体变异数之置信区间

(1) 当总体变异数 $\sigma^2$ 的点估计  $S^2$ 已经求出后

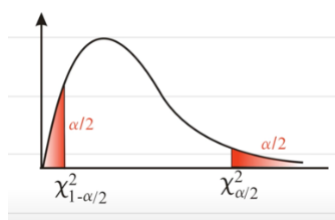
$$(2) \quad \frac{(n-1)s^2}{\sigma^2} \sim \chi^2$$

$$(3) \quad P(\chi^2_{1-\alpha/2} < \chi^2 < \chi^2_{\alpha/2}) = 1 - \alpha$$

$$(4) \quad P(\chi^2_{1-\alpha/2} < \frac{(n-1)s^2}{\sigma^2} < \chi^2_{\alpha/2}) = 1 - \alpha$$

$$(5) \quad \text{置信区间} P\left(\frac{(n-1)s^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}\right) = 1 - \alpha$$

$\alpha$



#### 泊松分布之置信区间

若 $x_1, x_2 \dots x_n$ 为取自泊松分布总体 poisson ( $\lambda$ ) 之一组随机样本且  $n > 30$ , 则参数 $\lambda$ 之 100(1- $\alpha$ )%信赖区间为

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

证明

(1)  $\lambda$ 的估计式为 $\bar{x}$

(2) 因为  $n > 30$ , 根据中央极限定理 $\bar{x} \sim N(\lambda, \frac{\lambda}{n})$

$$(3) \quad \text{根据 } Z = \frac{\bar{X} - \lambda}{\sqrt{\frac{\lambda}{n}}}$$

$$(4) \quad \lambda \text{ 的置信区间 } (\bar{X} - z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\lambda}{n}}, \bar{X} + z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\lambda}{n}})$$

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

当样本数超过 30 时, 可以使用中央极限定理, 套用  $Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$   $Z = \frac{\bar{X} - \lambda}{\sqrt{\frac{\lambda}{n}}}$

#### 估计误差与样本数

当  $x_1, x_2 \dots x_n$  为抽自 (1) 常态总体 (非常态总体但  $n > 30$ )

$$\bar{x} \sim N(\mu, \frac{\sigma^2}{n}) \quad \frac{d}{\sigma/\sqrt{n}} = Z_{\alpha/2}$$

$$P(|\bar{X} - \mu| \leq d) = 1 - \alpha \quad \text{误差: } d = Z_{\alpha/2} \cdot (\sigma/\sqrt{n})$$

$$P(|Z| \leq \frac{d}{\sigma/\sqrt{n}}) = 1 - \alpha \quad \text{样本数: } n = (\frac{Z_{\alpha/2}}{d})^2 \cdot \sigma^2$$

#### 总体比例估计时

当  $x_1, x_2 \dots x_n$  为抽自二项分布总体, 且  $n > 30$ , 则由  $\hat{p}$  估计 P 的误差估计

$$\text{误差: } d = z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}\hat{q}}{n}} \quad \text{样本数: } n = (\frac{Z_{\alpha/2}}{d})^2 \cdot \hat{p}\hat{q}$$

#### 第 21 章 双总体区间估计一

##### 双总体平均数之差的置信区间(大样本)

假设  $X_1$  和  $X_2$  为正态总体,  $\bar{x}_1 - \bar{x}_2$  抽样分布的标准差

分三种情况:

$$1. \quad \text{两总体变异数已知 } \sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$2. \quad \text{两总体变异数未知 } \sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$3. \quad \text{两总体变异数未知, 但确认相等 } \sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \text{ (加权的做法)}$$

#### 加权样本标准差

$$1. \quad \text{第一组样本: } s_1^2 = \frac{\sum_{i=1}^{n_1} (x_{1i} - \bar{x}_1)^2}{n_1 - 1}$$

$$2. \quad \text{第二组样本: } s_2^2 = \frac{\sum_{i=1}^{n_2} (x_{2i} - \bar{x}_2)^2}{n_2 - 1}$$

$$3. \quad \text{混合样本标准差 } s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

#### 置信区间:

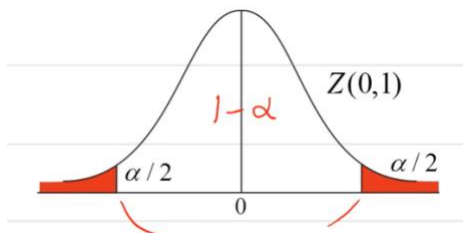
(1) 当两个总体变异数  $\sigma_1^2, \sigma_2^2$  已知时



$$\text{统计量 } z_{\alpha/2} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \quad z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{x}_1 - \bar{x}_2}} \quad \sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$P\left(-Z_{\alpha/2} < \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} < Z_{\alpha/2}\right) = 1 - \alpha$$

$$P\left[(\bar{X}_1 - \bar{X}_2) - Z_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{X}_1 - \bar{X}_2) + Z_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right] = 1 - \alpha$$



#### 双总体平均数之差的置信区间(小样本)

假设  $X_1$  和  $X_2$  为正态总体， $\bar{x}_1 - \bar{x}_2$  抽样分布的标准差

(2) 当两个总体变异数  $\sigma_1^2$ ,  $\sigma_2^2$  未知，但确认相等

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \quad \text{其中混合样本方差 } s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$\text{置信区间: } P\left[(\bar{X}_1 - \bar{X}_2) - t_{\alpha/2} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \mu_1 - \mu_2 < (\bar{X}_1 - \bar{X}_2) + t_{\alpha/2} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right] = 1 - \alpha$$

(3) 当两个总体变异数  $\sigma_1^2$ ,  $\sigma_2^2$  未知，确认不相等，以  $S_1$ ,  $S_2$  取代  $\sigma_1$ ,  $\sigma_2$ ，用 t 分布修正

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_{\bar{x}_1 - \bar{x}_2}} \quad s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

$$\text{置信区间: } P\left[(\bar{X}_1 - \bar{X}_2) - t_{\alpha/2} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{X}_1 - \bar{X}_2) + t_{\alpha/2} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right] = 1 - \alpha$$

#### 双总体平均数之差的置信区间(相依样本)

eg: 减肥药，有 6 人进行疗程，并将前后体重记录如下，试求前后差之 95% 的置信区间

	$X_1$	$X_2$	D	$D^2$
1	85	81	4	16
2	93	86	7	49
3	72	64	8	64
4	71	63	8	64
5	96	88	8	64
6	88	87	1	1

			36	258
--	--	--	----	-----

相依（成对）样本处理:

(1) 令  $\bar{D}_i = x_{1i} - x_{2i}$  视为一个样本处理

(2) 使用统计量

(A) 大样本总体变异数已知  $Z = \frac{\bar{D} - \mu_D}{\sigma_D}$  (C) 小样本总体常态变异数已知  $Z = \frac{\bar{D} - \mu_D}{\sigma_D}$

(B) 大样本总体变异数未知  $Z = \frac{\bar{D} - \mu_D}{s_D}$  (D) 小样本总体常态变异数未知  $t = \frac{\bar{D} - \mu_D}{s_D}$

(3) 大样本、小样本总体变异数已知  $(\bar{D} - Z_{\alpha/2} \cdot \sigma_D, \bar{D} + Z_{\alpha/2} \cdot \sigma_D)$

大样本总体变异数未知  $(\bar{D} - Z_{\alpha/2} \cdot s_D, \bar{D} + Z_{\alpha/2} \cdot s_D)$

小样本总体常态变异数未知  $(\bar{D} - t_{\alpha/2} \cdot s_D, \bar{D} + t_{\alpha/2} \cdot s_D)$

## 第 22 章 双总体区间估计二

### 两个总体比例之差的置信区间（大样本）

1. 当两个总体比例  $P_1$  和  $P_2$  的点估计  $\hat{p}_1$  和  $\hat{p}_2$  已经求出后
2. 根据中央极限定理 (CLT),  $\hat{p}_1 - \hat{p}_2$  的抽样分布  $\hat{p}_1 - \hat{p}_2 \sim N(P_1 - P_2, \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2})$
3. 检验统计量  $Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{s_{\hat{p}_1 - \hat{p}_2}}$
4. 置信区间  $\left( (\hat{p}_1 - \hat{p}_2) - Z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}, (\hat{p}_1 - \hat{p}_2) + Z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} \right)$

### 两个总体变异数比率的置信区间

1. 两个总体变异数之比  $\frac{\sigma_1^2}{\sigma_2^2}$  很自然以估计  $\frac{s_1^2}{s_2^2}$
2.  $\frac{s_1^2/s_2^2}{\sigma_1^2/\sigma_2^2} \sim F$
3.  $P(f_{1-\alpha/2}(v_1, v_2) < F < f_{\alpha/2}(v_1, v_2)) = 1 - \alpha$
4.  $P(f_{1-\alpha/2}(v_1, v_2) < \frac{s_1^2/s_2^2}{\sigma_1^2/\sigma_2^2} < f_{\alpha/2}(v_1, v_2)) = 1 - \alpha$
5. 置信区间  $P\left(\frac{s_1^2}{s_2^2} \frac{1}{f_{\alpha/2}(v_1, v_2)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2} \frac{1}{f_{1-\alpha/2}(v_1, v_2)}\right) = 1 - \alpha$

### 样本数的选择

目的:

1. 样本数提高, 误差会降低
2. 实验设计需要规划所需样本

方法:

设定最大容忍误差  $\Delta$ , 若总体  $\sigma$  已知, 考虑一个单尾假设检验问题:

$$H_0: \mu = \mu_0$$

$$H_1: \mu > \mu_0$$

最少样本公式:

设定最大容忍误差 $\Delta$ ，， 我可以先假设 $\mu = \mu_0 + \Delta$ ，在显著水平 $\alpha$ 之下

$$1 - \beta = P\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} > Z_\alpha \mid H_1 \text{ is true}\right)$$

$$1 - \beta = P\left(Z > Z_\alpha - \frac{\Delta}{\frac{\sigma}{\sqrt{n}}}\right) - Z_\beta = Z_\alpha - \frac{\Delta}{\frac{\sigma}{\sqrt{n}}}$$

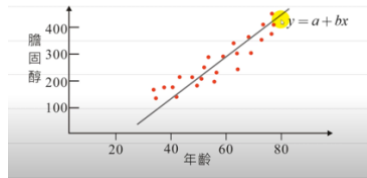
$$\frac{\Delta}{\sigma/\sqrt{n}} = (Z_\alpha + Z_\beta)$$

$$n = \frac{(Z_\alpha + Z_\beta)^2 \sigma^2}{\Delta^2}$$

$$\text{若考慮雙尾 } n = \frac{(Z_{\alpha/2} + Z_\beta)^2 \sigma^2}{\Delta^2}$$

### 第 23 章 回归一

目的：探索两个或两个以上变数之间的关系。希望可以找到一个公式可以预测变数



设  $y$  为随机变数，自变数  $x$  假设为已知之固定常数，若  $y$  與  $x$  具有直线回归关系式。为了方便我们定义  $Y \mid x$

为对于固定  $x$  随机变数  $Y, Y, f(y|x)$  为概率分布，总体回归直线

$$E(Y|x) = \alpha + \beta x$$

对  $y$  的估计量为  $y$  的估计量为  $\alpha$ ， $\beta$  的估计量为  $b$  样本回归直线

$$y = a + bx$$

总体回归方程式

$$y_i = \alpha + \beta x_i + e_i$$

$$E(y_i) = \alpha + \beta x_i$$

$$E(e_i) = 0$$

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

$$e_i \sim N(0, \sigma^2)$$

$$\text{Cov}(e_i, e_j) = 0$$

样本回归方程式

$$y_i = a + bx_i + \varepsilon_i$$

$$\hat{y}_i = a + bx_i$$

$$\sum_{i=1}^n \varepsilon_i = 0$$

$$\sum_{i=1}^n \varepsilon_i^2 \text{ 最小}$$

### 最小平方方法

依据：

1. 残差总和为 0
2. 残差平方和最小
3. 残差与变数之间的共变数为 0

回归系数 $\alpha$ 和  $\beta$  的估计

$$SSE = \sum_{i=1}^n (y_i - \hat{y})^2 \text{极小}$$

$$\sum_{i=1}^n (y_i - a - bx_i)^2 \text{极小}$$

符号记忆公式

$$\bar{x} = \frac{\sum x_i}{n} \quad \bar{y} = \frac{\sum y_i}{n}$$

$$S_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n} \quad x_i \text{ 之变异数}$$

$$S_{yy} = \sum (y_i - \bar{y})^2 = \sum y_i^2 - \frac{(\sum y_i)^2}{n} \quad y_i \text{ 之变异数}$$

$$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n} \quad x_i \text{ 與 } y_i \text{ 之共变异数}$$

$$b = \frac{\sum x_i y_i - (\sum x_i \sum y_i)/n}{\sum x_i^2 - (\sum x_i)^2/n} = \frac{S_{xy}}{S_{xx}} \quad a = \bar{y} - b\bar{x}$$

判定系数

评估回归方程的适配度，评估回归方程的解释能力

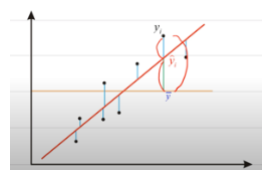
$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad \text{随机变异}$$

$$SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \quad \text{回归变异}$$

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2 \quad \text{总变异}$$

$$(y_i - \bar{y}) = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

$$SST = SSE + SSR$$



判定系数 (coefficient of determination) 即是回归平方 SSR 和占 SST 的比例，常用  $R^2$  表示，范围[0,1]，越靠近 1 的回归方程的适配度越高

$$1 = \frac{SSE}{SST} + \frac{SSR}{SST} \quad R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

回归参数的估计

利用判定系数  $R^2$  (coefficient of determination) 估计回归方程式之配适度 (goodness of fit)，其必须先经过显著性检验。当未达到显著性水准时，判定数  $R^2$  数值高低不具任何意义

基本假设

回归分析之显著性检定必须依据下列误差项  $\epsilon_i$  的假设条件

- (1) 各观测点之误差项  $\epsilon_i$  的平均值或期值为 0 ( $\sum_{i=1}^n \epsilon_i = 0$ )
- (2) 各观测点之误差项  $\epsilon_i$  之间相互独立. ( $Cov(\epsilon_i, \epsilon_j) = 0, i \neq j$ )
- (3) 各观测点之误差项  $\epsilon_i$  之变异数  $\sigma^2$  皆相等  $Var(\epsilon_i) = \sigma^2, i = 1 \dots n$  也就是说残差项之变异数具有齐一性
- (4) 各观测点之误差项  $\epsilon_i$  属于常态分布

## 变异数估计

因变量 $y_i$ 以自变量 $x_i$ 建立的回归模型  $\hat{y}_i = a + bx_i + \epsilon_i$

误差项 $\epsilon_i$ 的变异数即是因变量 $y_i$ 在回归模型中的变异数 $\sigma^2$

$$\sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (y_i - \hat{y})^2 = \sum_{i=1}^n (y_i - a - bx_i)^2 = SSE$$

$$\hat{\sigma} = \frac{SSE}{df} = \frac{SSE}{n-2} \text{ 為 } \sigma^2 \text{ 的估計量}$$

## 回归系数 b 的期望值

$$b = \frac{S_{xy}}{S_{xx}}$$

$$E(y_i) = \alpha + \beta x_i$$

$$E(b) = \beta$$

回归系数 b 的变异数

$$Var(b) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2} = \frac{\sigma^2}{S_{xx}}$$

计算回归系数 a 与 b

$$b = \frac{S_{xy}}{S_{xx}} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$$

$$a = \bar{y} - b\bar{x}$$

$$E(b) = \beta$$

$$E(a) = \alpha$$

$$Var(b) = \sigma_b^2 = \frac{\sigma^2}{S_{xx}}$$

$$Var(a) = \sigma_a^2 = \frac{\sum x_i^2}{nS_{xx}} \sigma^2$$

## 回归系数 b 的抽样分布

回归系数  $b \sim N(\beta, \sigma_b^2)$  其中  $\sigma_b^2 = \frac{\sigma^2}{S_{xx}}$

(1) 一般来说  $\sigma$  都是未知, 故以  $\hat{\sigma}$  取代  $\sigma$   $\hat{\sigma} = \frac{SSE}{n-2}$

(2) 当为大样本时  $\frac{b-\beta}{\sigma_b} \sim Z = N(0,1)$

(3) 当为小样本时  $\frac{b-\beta}{\sigma_b} \sim t = t(n-2)$

## 第 24 章 回归二

### 显著性检验

利用判定数  $R^2$  (coefficient of determination) 估计回归方程式之配适度 (goodness of fit), 其必须先

经过着性检定。当未达到显著性水准时, 判定系数  $R^2$  数值高低不具任何意义

显著性检验: (1) 检定模型 (2) 检定系数

### 基本假设

回归分析之显著性检定必须依据下列误差项 $\epsilon_i$ 的假设条件

(1) 各观测点之误差项 $\epsilon_i$ 的平均值或期值为 0 ( $\sum_{i=1}^n \epsilon_i = 0$ )

(2) 各观测点之误差项 $\epsilon_i$ 之间相互独立. ( $Cov(\epsilon_i, \epsilon_j) = 0, i \neq j$ )

(3) 各观测点之误差项 $\epsilon_i$ 之变异数 $\sigma^2$ 皆相等  $Var(\epsilon_i) = \sigma^2, i = 1 \dots n$ 也就是说残差项之异数具有齐一性

(4) 各观测点之误差项 $\epsilon_i$ 属于常态分布

ANOVA 显著性检验

(1) 建立 ANOVA 表

(2)  $H_0$ : 回归方程无解释能力

$H_1$ : 回归方程有解释能力

(3) 求出  $\frac{SSR/1}{SSE/n-2} = \frac{MSR}{MSE} = F$

$F > F_{\alpha}(1, n-2)$  则拒绝  $H_0$

变异	SS	自由度	均方	F 值
回归	SSR=	1	MSR=	$\frac{MSR}{MSE} =$
误差	SSE=	n-2	MSE=	
综合	SST=	n-1		

多元回归

当自变量不止一个时的回归分析方法

(1) 简单回归  $y = a + bx$

(2) 二元回归  $y = a + b_1x_1 + b_2x_2$

(3) 多元回归  $y = a + b_1x_1 + b_2x_2 + \dots + b_kx_k$

多元回归的原则

1. 预测变量间的相关系数越低越好，表示独立；预测项  $X_i$  与响应量  $Y_i$  的相关度越高越好
2. 每个预测变量间的相关系数为 0，可避免多元共线性的问题；变量间的相关系数若高于 0.8，将彼此相关系数较高的预测变项只取一个重要变项

$$y_i = a + b_1x_{1,i} + b_2x_{2,i} + \dots + b_kx_{k,i}$$

流程：

1. 找出回归系数
2. 估计误差的标准误
3. 估计回归系数的标准误（信赖区间）
4. 检验回归系数的显著性
5. 以判定系数  $R^2$  判断模型的适配度

使用 F 检验

(1) 虚无假设  $H_0: \beta_1 = \beta_2 = \dots \beta_k = 0$

对立假设  $H_1: \beta_1 = \beta_2 = \dots \beta_k = 0$  不全为 0

(2) 建立 ANOVA 表

		自由度	均方	F 值	F 临界值
回归 SSR	$SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$	k	$MSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 / K$	$\frac{MSR}{MSE}$	$F_{\alpha}(k, n-k-1)$

误差 SSE	$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$	n-k-1	$MSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2 / (n - k - 1)$		
综合 SST	$SST = \sum_{i=1}^n (y_i - \bar{y})^2$	n-1			

## 第 25 章 相关

协方差 (covariance)

设 X 和 Y 为随机变数设,  $E(X) = \mu_1$ ,  $E(Y) = \mu_2$ , X 和 Y 的协方差为:

$$\text{Cov}(X, Y) = E[(X - \mu^1)(Y - \mu^2)] = E(XY) - E(X)E(Y)$$

皮尔逊相关系数

$$\text{总体相关系数 } \rho = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{E(X - \mu_X)(Y - \mu_Y)}{\sigma_X \sigma_Y}$$

$$\text{样本相关系数 } r = \frac{\sum_{i=1}^n (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum_{i=1}^n (X - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y - \bar{Y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}} \sqrt{S_{yy}}}$$

## 相关系数的检验

- (1) 采用皮尔逊相关系数, 主要是测量两连续变量关系的强弱
- (2) 建立假设  $H_0: \rho = 0$  对立假设  $H_1: \rho \neq 0$
- (3) 选择显著水平  $\alpha$
- (4) 采用 t 检验, 统计量  $t = r \sqrt{\frac{n-2}{1-r^2}}$  检验相关系数是否为 0
- (5) 在回归分析中, 检验统计量  $t = \frac{b}{\sigma_b}$  检验回归系数  $\beta$  是否为 0

注意:

1.  $\rho = 0$  时, 使用 t 检验, 检验统计量与检验回归斜率 b 等价
2.  $\rho \neq 0$  时, 使用 fisher 转换后用 z 检验

## 相关与回归

命题	相关系数	简单线性回归
数据尺度	X 与 Y 均为连续性	X 与 Y 均为连续性
假设 X 与 Y 为直线关系	是	是
指定相应变量和自变量	否	Y 为因变量, X 是自变量
因果关系	不一定	不一定
正负	相同	相同公式
公式	$r = \frac{S_{xy}}{\sqrt{S_{xx}} \sqrt{S_{yy}}}$	$b = \frac{S_{xy}}{S_{xx}}$

$H_0$	$\rho = 0$	$\beta=0$
检验方法	T 检验	T 或 F 检验
检验统计量	$t = r \sqrt{\frac{n-2}{1-r^2}}$ df = n-2	$t = \frac{b}{\sigma_b}$ , df = n-2

## 第 26 章 卡方检验

### 卡方检验的观念

统计资料的分类：

- 1.数量资料：母体以平均数、变异数为主。检验方法以母数统计/参数统计为主
- 2.类别资料：总体参数以比例为主，如满意度，李克特 5 等级满意度意见调查。检验方法以非参数估计为主/无母数统计。

卡方检验的观念与方法：

- 1.依据资料观察值 (O) 次数记录于表格中的储存格
- 2.依据条件或要求计算出每个储存格的期望值(E)
- 3.利用 $\sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i} \sim \chi^2$ 的观念，若期望值(E)与观察值 (O) 的差距不大，则卡方值很小，**得到不显著的结果**
- 4.每个储存格数量不可低于 5，若低于 5 要与旁边的储存格合并

### 适合度检验

目的：

利用样本资料检验总体是否符合某种特定分布

观念与方法：

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

### 独立性检验

目的：

当观察值具有两种特征时使用二因子列联表 (two factor contingency table) 进而检定两因子是否相关 · 二因子列联表习惯采用列 (row) 与 c 行 (column) 又称为 r×c 表

自由度  $rc - r - c + 1 = (r-1)(c-1)$

### 齐一性检验

检验两个或两个以上的总体的某一特性的分布是否相同或相近

自由度  $rc - r - c + 1 = (r-1)(c-1)$



## 第 27 章 方差分析

### 方差分析的观念

目的：

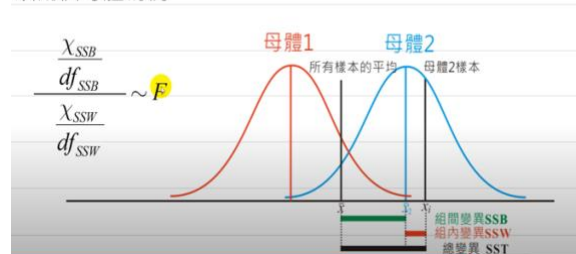
1. 检验两个总体 $\mu$ 是否有差异经常采用 t 检验
2. 检验 3 个或 3 个以上总体 $\mu$ 是否相当的统计方法，称为方差分析。

例子：

1. 来自不同区域学生英文成绩的差别
2. 多家饭店满意度的比较
3. 不同民族身高的比较
4. 施加不同肥料对作物产量的影响

原理：利用组间变异与各组组内变异各自处以其自由度服从 F 分布的性质。

以兩個母體為例



$$\frac{\frac{\chi_{SSB}}{df_{SSB}}}{\frac{\chi_{SSW}}{df_{SSW}}} \sim F$$

### 单因子方差分析

需要条件：

- 常态性：总体要正态分布
- 独立性：样本是独立的简单随机抽样
- 同质性：方差要求同质

定义：单因子方差分析的资料是来自 k 个总体的每个总体的简单随机样本，而由第 j 个总体中抽出 $n_j$ 个观察值

(1)  $x_{(1,1)}, x_{(1,1)} \dots x_{(n_1,1)}$  第一个总体，有  $n_1$  个观察值，总体平均 $\mu_1$

(2)  $x_{(1,2)}, x_{(1,2)} \dots x_{(n_2,2)}$  第 2 个总体，有  $n_2$  个观察值，总体平均 $\mu_2$

$$x_{x,j} = u_j + \epsilon_{i,j} \quad \epsilon_{i,j} \sim N(0, \sigma^2)$$

变异之间关系

$$\text{总平均数: } \bar{\bar{x}} = \frac{1}{N} \sum_{j=1}^k \sum_{i=1}^{n_j} X_{ij} \quad N = \sum_{j=1}^k n_j$$

$$\text{第 } j \text{ 个总体的平均数: } \bar{x}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} X_{ij}$$

$\sum \sum (X_{ij} - \bar{\bar{x}})^2$	$\sum \sum (\bar{x}_j - \bar{\bar{x}})^2$	$\sum \sum (X_{ij} - \bar{x}_j)^2$
总变异 sst	组间 ssb	组内 ssw
N-1	k-1	N-k

来源/平方和	自由度	均方和	F 值
组间 SSB	k-1	$MSB = \frac{SSB}{(K-1)}$	$F = \frac{MSB}{MSW}$
组内 SSW	N-k	$MSW = \frac{SSW}{(N-K)}$	
总和 SST	N-k		

### Bonferroni 事后比较

平均数的多重比较

若  $H_0: \mu_1 = \mu_2 \dots = \mu_n$  在 F 检验遭受拒绝，表示并不是每个总体平均数都相等，此时应进行事后比较观念：

使用置信区间的观念，若有 n 个总体，需进行  $C_2^n$  组检验

在小样本中，总体为常态且方差未知时  $\bar{x}_i - \bar{x}_j$  的抽样分布为 t 分布， $u_i - u_j$  的  $(1-\alpha)\%$  置信区间为

$$\left[ (\bar{X}_i - \bar{X}_j) - t_{\alpha/2} \cdot \sqrt{\frac{1}{n_i} + \frac{1}{n_j}} < \mu_1 - \mu_2 < (\bar{X}_i - \bar{X}_j) + t_{\alpha/2} \cdot \sqrt{\frac{1}{n_i} + \frac{1}{n_j}} \right]$$

加权标准差

两个总体 X 与 Y 平均数比较时

- (1) 若标准差已知为  $\sigma_1$  和  $\sigma_2$

$$Var(\bar{x} - \bar{y}) = Var(\bar{x}) + Var(\bar{y}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

- (2) 若标准差未知

$$Var(\bar{x} - \bar{y}) = Var(\bar{x}) + Var(\bar{y}) = \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}$$

- (3) 若标准差未知，但相等

$$(\bar{x} - \bar{y}) = Var(\bar{x}) + Var(\bar{y}) = \frac{s^2}{n_1} + \frac{s^2}{n_2} = s^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)$$

$$\sigma_{\bar{x}-\bar{y}} = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad s_p = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}$$

BONFERRONI 法

- (1) 置信区间

$$(\bar{X}_i - \bar{X}_j) \pm t_{\alpha/2} \cdot s_p \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

(2)  $s_p$  使用  $\sqrt{MSE}$   $MSE = \frac{SSE}{df}$

(3) 若有  $n$  个总体需要交互比较, 置信水平  $(1-\alpha)$  改为  $1 - \alpha/c_2^n$

联立置信区间:

若  $E_i I = 1 \dots m$  表个置信区间显著水平,  $P(E_1 \cap \dots \cap E_m) = P(E_i)^m$

$P(E_i)^m = 1 - \alpha$   $P(E_1) = (1 - \alpha)^{1/m} \approx 1 - \alpha/m$

POST-HOC 多重比较

$\bar{x}_1 = 5.2$	$\bar{x}_2 = 7.6$	$\bar{x}_3 = 5$	$MSE = 2.167$ $df = 12$	$(\bar{X}_i - \bar{X}_j) \pm t_{\alpha/2} \cdot s_p \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$
				$t_{0.05/3}(12) = 2.78$
$H_0: \mu_1 = \mu_2$	$2.4 \pm (2.78) \sqrt{2.167} \sqrt{\frac{1}{5} + \frac{1}{5}} = 2.4 \pm 2.58$			$(-0.08, 4.98)$
$H_0: \mu_1 = \mu_3$	$0.2 \pm (2.78) \sqrt{2.167} \sqrt{\frac{1}{5} + \frac{1}{5}} = 0.2 \pm 2.58$			$(-2.38, 2.78)$
$H_0: \mu_2 = \mu_3$	$2.6 \pm (2.78) \sqrt{2.167} \sqrt{\frac{1}{5} + \frac{1}{5}} = 2.6 \pm 2.58$			$(0.02, 5.18)$

$\mu_2 > \mu_3$

SHEFFE 事后比较

若  $H_0: \mu_1 = \mu_2 \dots = \mu_n$  在 F 检验遭受拒绝, 表示并不是每个总体平均数都相等, 此时应进行事后比较观念:

使用置信区间的观念, 若有  $n$  个总体, 需进行  $C_2^n$  组检验

在小样本中, 总体为常态且方差未知时  $\bar{x}_i - \bar{x}_j$  的抽样分布为 t 分布,  $u_i - u_j$  的  $(1-\alpha)\%$  置信区间为

$$(\bar{X}_i - \bar{X}_j) \pm t_{\alpha/2} \cdot s_p \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

加权标准差

两个总体 X 与 Y 平均数比较时

(1) 若标准差已知为  $\sigma_1$  和  $\sigma_2$

$$Var(\bar{x} - \bar{y}) = Var(\bar{x}) + Var(\bar{y}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

(2) 若标准差未知

$$Var(\bar{x} - \bar{y}) = Var(\bar{x}) + Var(\bar{y}) = \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}$$

(3) 若标准差未知, 但相等

$$(\bar{x} - \bar{y}) = Var(\bar{x}) + Var(\bar{y}) = \frac{s^2}{n_1} + \frac{s^2}{n_2} = s^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)$$

$$\sigma_{\bar{x}-\bar{y}} = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad s_p = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$$

SHEFFE 法

(1) 置信区间

$$(\bar{X}_i - \bar{X}_j) \pm t_{\alpha/2} \cdot s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

(2)  $s_p$  使用  $\sqrt{MSE}$      $MSE = \frac{SSE}{df}$

(3) SHEFFE 法是利用  $t^2 = F$  的观念，将  $t_{\alpha/2}$  改为  $\sqrt{((k-1)F_{\alpha}(k-1, n-k))}$

(4) SHEFFE 法的置信区间为

$$(\bar{X}_i - \bar{X}_j) \pm \sqrt{((k-1)F_{\alpha}(k-1, n-k))} \cdot s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

POST-HOC 多重比较同 BONFERRONI 法，置信区间用  $(\bar{X}_i - \bar{X}_j) \pm \sqrt{((k-1)F_{\alpha}(k-1, n-k))} \cdot s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

第 28 章 双因子方差分析

双因子方差分析（无交互）

实验设计：

目的：控制不是研究对象的因子，以获得要研究因子的影响效果

		因子A			
		處理1	處理2	處理3	處理4
因子B	處理1				
	處理2				
	處理3				

样本资料

处理 TREATMENT(A)						
区组 BLOCK (B)		1	2	...	a	列总和
	1	$X_{11}$	$X_{12}$		$X_{1a}$	$B_1 = \sum_{j=1}^a X_{1j}$
	2	$X_{21}$	$X_{22}$		$X_{2a}$	$B_2 = \sum_{j=1}^a X_{2j}$
	⋮	⋮	⋮	...		

	b	$X_{b1}$	$X_{b2}$	...	$X_{ba}$	$B_b = \sum_{j=1}^a X_{bj}$
	行总和	$T_1 = \sum_{i=1}^b X_{i1}$	$T_2 = \sum_{i=1}^b X_{i2}$			$S = \sum_{j=1}^a \sum_{i=1}^b X_{ij}$

总差异 = 处理方式差异 + 区集差异 + 残差

$$X_{ij} - \bar{X} = \left( \bar{T}_j - \bar{X} \right) + \left( \bar{B}_i - \bar{X} \right) + \left( X_{ij} - \bar{T}_j - \bar{B}_i + \bar{X} \right)$$

两边取平方，取总和，可以证明

$$\sum_{j=1}^a \sum_{i=1}^b \left( X_{ij} - \bar{X} \right)^2 = b \cdot \sum_{j=1}^a \left( \bar{T}_j - \bar{X} \right)^2 + a \cdot \sum_{i=1}^b \left( \bar{B}_i - \bar{X} \right)^2 + \sum_{j=1}^a \sum_{i=1}^b \left( X_{ij} - \bar{T}_j - \bar{B}_i + \bar{X} \right)^2$$

$$SST = SSA + SSB + SSE$$

双因子无交互 ANOVA 表

来源	平方和	自由度	均方	F 值
处理方式	SSA	a-1	$MSA = \frac{SSA}{(a-1)}$	$F_A = \frac{MSA}{MSE}$
区集	SSB	b-1	$MSB = \frac{SSB}{(b-1)}$	$F_B = \frac{MSB}{MSW}$
残差	SSE	(a-1)(b-1)	$MSE = \frac{SSE}{(a-1)(b-1)}$	
总和	SST	ab-1		

双因子方差分析（有交互）

因变量同时受到 A 与 B 因子影响，若考虑 AB 间交叉影响，每个黄色区块应施予 2 个以上的本（重复试验），每个区块样本数可以不同，若相同处理较容易（均衡实验设计）

		因子A			
		處理1	處理2	處理3	處理4
因子B	處理1				
	處理2				
	處理3				

重复实验样本资料

因子(A)						
因子 (B)		1	2	...	a	列总和
	1	$X_{111}, X_{112} \dots X_{11n}$	$X_{121}, X_{122} \dots X_{12n}$		$X_{1a1}, X_{1a2} \dots X_{1an}$	$B_1 = \sum_{j=1}^a \sum_{i=1}^n X_{1jk}$
	2	$X_{211}, X_{212} \dots X_{21n}$	$X_{221}, X_{222} \dots X_{22n}$		$X_{2a1}, X_{2a2} \dots X_{2an}$	$B_2 = \sum_{j=1}^a \sum_{i=1}^n X_{2jk}$
	⋮	⋮	⋮	...		
	b	$X_{b11}, X_{b12} \dots X_{b1n}$	$X_{b21}, X_{b22} \dots X_{b2n}$	...	$X_{ba1}, X_{ba2} \dots X_{ban}$	$B_b = \sum_{j=1}^a \sum_{i=1}^n X_{bjk}$
	行总和	$T_1 = \sum_{j=1}^b \sum_{k=1}^n X_{i1k}$	$T_2 = \sum_{j=1}^b \sum_{i=1}^n X_{i2k}$		$T_a = \sum_{j=1}^a \sum_{i=1}^n X_{iak}$	$S = \sum_{j=1}^a \sum_{i=1}^b \sum_{k=1}^n X_{ijk}$

总变异 = A 因子变异 + 因子变异 + 交互作用变异 + 残差变异

$$SST = SSA + SSB + SSAB + SSE$$

$abn-1$   $a-1$   $b-1$   $(a-1)(b-1)$   $(abn-ab)$  -自由度

双因子有交互的 ANOVA 表

来源	平方和	自由度	均方	F 值
因子 A	SSA	$a-1$	$MSA = \frac{SSA}{(a-1)}$	$F_A = \frac{MSA}{MSE}$
因子 B	SSB	$b-1$	$MSB = \frac{SSB}{(b-1)}$	$F_B = \frac{MSB}{MSW}$
因子 AB	SSAB	$(a-1)(b-1)$	$MSAB = \frac{SSAB}{(a-1)(b-1)}$	$F_{AB} = \frac{MSAB}{MSE}$
残差	SSE	$ab(n-1)$	$MSE = \frac{SSE}{ab(n-1)}$	
总和	SST	$abn-1$		

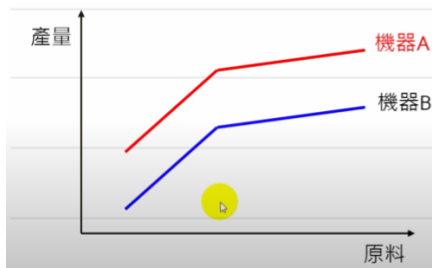
分析结果注意事项

1. 若检验交互作用的结果不显著，则主效应的检验教效力，即具有解释性
2. 若检验交互作用的结果显著（即有交互效应），表示两因子联合对应变数产生效应，因此单独执行主效应检验就变得无意义
3. 有交互效应时应以其他方法进行分析，如：固定某一因子去分析另一因子的效应
- 4.

交互作用几何意义

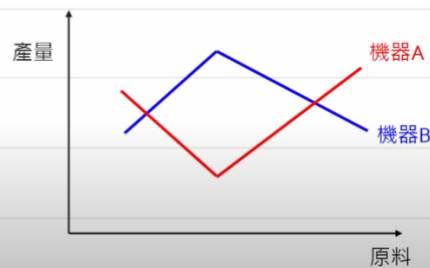
## 無交互作用

原料改變而造成產量改變量不會受機器因子的影響



## 有交互作用

原料改變而造成產量改變量會受機器因子的影響



## 第 29 章 无母数检定/非参检验—

### 无母数检定

非参检验特性：

1. 总体数值分布的形态不稳定
2. 统计推论的对象不局限于特定的总体参数
3. 总参参数统计的假设条件参数太少

非参检验缺点：

1. 总体分布已知时，非参数估计效果差
2. 缺乏相关表格可以用
3. 针对常态分布资料检验时，会使检定力降低

### 统计方法

符号检验(sign test)：

1. 单一总体中位数的假设
2. 两总体的中位数是否相同的假设检验

Wilcoxon rank sum test

Mann-Whitney U test

Kruskal- Wallis test(K-W test)

Friedman test

Spearman test 相关性检验

Kolmogorov-Smirnov test(K-S test)

### 符号检验—单一总体中位数检验

方法：

- (1) 先讲观察值减去 $\eta_0$  ( $\eta$ )  $d_i = X_i - \eta_0$
- (2) 令 X 表  $d_i$  “+”的个数
- (3) 检验 (以 p 值法)

$$\begin{array}{lll} H_0: \eta = \eta_0 & H_0: \eta = \eta_0 & H_0: \eta = \eta_0 \\ H_1: \eta \neq \eta_0 & H_1: \eta > \eta_0 & H_1: \eta < \eta_0 \end{array}$$

$$P = 2 \cdot \min \begin{cases} P(X \leq x | p = 0.5) \\ P(X \geq x | p = 0.5) \end{cases} \quad P(X \geq x | p = 0.5) \quad P(X \leq x | p = 0.5)$$

注意：

符号检定 (Sign Test) 是用来

- (1) 检定母体中位数是否等于某特定值
- (2) 可以用来检定两组母体的中位数是否相等
- (3) 以正号、负号的个数做为检定的基础
- (4) S+的抽样分布是二项分配

#### 符号检验—中位数的区间估计

假设  $X_1 + X_2 \dots + X_n$  为取自总体的一组随机样本

$$X_1, X_2 \dots X_r \dots \dots \dots X_s \dots, X_{n-1}, X_n$$

$$P(X_r < \eta < X_s)$$

$$\sum_{i=r}^s C_i^n \left(\frac{1}{2}\right)^n = 1 - \alpha$$

则称  $(X_r, X_s)$  为中位数  $\eta$  的  $(1 - \alpha)\%$  置信区间

#### 符号检验—两个相关总体中位数的区间估计

方法：

- (1) 先讲观察值减去  $\eta_0$  ( $\eta$ )  $d_i = X_i - \eta_0$
- (2) 令 X 表  $d_i$  “+” 的个数
- (3) 检验 (以 p 值法)

$$\begin{array}{lll} H_0: \eta = \eta_0 & H_0: \eta = \eta_0 & H_0: \eta = \eta_0 \\ H_1: \eta \neq \eta_0 & H_1: \eta > \eta_0 & H_1: \eta < \eta_0 \end{array}$$

$$P = 2 \cdot \min \begin{cases} P(X \leq x | p = 0.5) \\ P(X \geq x | p = 0.5) \end{cases} \quad P(X \geq x | p = 0.5) \quad P(X \leq x | p = 0.5)$$

#### WILCOXON--两独立母体中位数检验

WILCOXON RANK SUM TEST

1. 针对两个独立母体检验中位数
2. 与母体独立样本 t 检验对应

要求：

1. 两总体形状与方差相同
2. 资料要求的顺序尺度即可

理论基础：



当两总体一致时，两组样本观察值的等级和非常接近；否则两等级和相去甚远，表示两总体不可能一致，所以要拒绝  $H_0$

检验方法：

设第一个总体：中位数  $\eta_1$   $X_{11}, X_{12} \dots X_{1n_1}$

设第二个总体：中位数  $\eta_2$   $X_{21}, X_{22} \dots X_{2n_2}$

混合再排序，并记录等级，遇到同分以平均等级代替，定义  $R_i = \text{Rank of } X_{ij}$

$W_1 =$  第一个样本等级和  $W_1 + W_2 = 1+2+\dots+n = \frac{n(n+1)}{2}$

(1) 小样本时查 WILCOXON 表

(2) 大样本时以常态近似。 $W_1 \sim N(\frac{n_1(n_1+n_2+1)}{2}, \frac{n_1n_2(n_1+n_2+1)}{12})$

重点整理

- (a) 两独立母体之比较
- (b) 只要其中一组样本大于 10 就称为大样本
- (c) 为样本数较小的样本数，统计量为等级和

(d)  $\mu_w = \frac{n_1(n_1+1)}{2} \quad \sigma_w = \sqrt{\frac{n_1n_2(n_1+1)}{12}}$

#### MANN WHITNEY U -- 两独立总体中位数检验

- (1) 与 Wilcoxon rank sum test 的相同，对两独立母体检定中位数
- (3) 做法开始与 Wilcoxon rank sum test 类似，差在统计量不同
- (4) 与母数独立样本检定对应

做法：

混合后定等级

$W_1 =$  第一个样本等级和

$W_2 =$  第二个样本等级和

$$U_1 = n_1n_2 + \frac{n_1(n_1+1)}{2} - W_1$$

$$U_2 = n_1n_2 + \frac{n_1(n_1+1)}{2} - W_2$$

$$U = \min(U_1, U_2)$$

检验方法

- (1) 小样本时查 WILCOXON U 表
- (2) 大样本时以常态近似。（只要其中一组样本大于 10）

$$U \sim N(\frac{n_1n_2}{2}, \frac{n_1n_2(n_1+1)}{12})$$

$$Z = \frac{U - \frac{n_1n_2}{2}}{\sqrt{\frac{n_1n_2(n_1+1)}{12}}} \sim N(0, 1)$$

## 第 30 章 无母数检定/非参检验二

### SPEARMAN 等级相关系数

定义：X 与 Y 为两组变数，其 SPEARMAN 等级相关系数

$$r_s = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}$$

其中  $d_i$  为  $X_i$  与  $Y_i$  的等级差，n 为资料组数

性质： $-1 \leq r_s \leq 1$

注意：

- (1) Spearman 等级相对数据条件的要求没有 Pearson 相数严格，只要两个变数的观测值是成对的等级资料或者是由连续数观测资料转化得到的等级资料。不论两数的分布形态、样本大小如何都可以 Spearman 等级相来进行研究
- (2) 计算  $r_s$  所得的检定力较 r 为低。若项符合之假设前提，建议尽可能使用 r (Pearson 相关系数)

### SPEARMAN 等级相关系数检验

- (1) 小样本，当  $n \leq 30$  时，查 Spearman 等级相关系数临界值
- (2) 大样本，当  $n > 30$  时，采用 Z 检验  $r_s = \pm \frac{Z}{\sqrt{n-1}}$

### KRUSKAL-WALLIS 检验

1. 当常态总体条件成立，方差相同，进行多个总体平均数检验，采用 ANOVA 方法
2. 以上条件不成立时，采用非参数检验 (KRUSKAL-WALLIS 检验)
3. 简称 K-W 检验
4. 可视为 WILCOXON RANK SUM TEST 由两总体检验延伸至多个总体

与 ANOVA 比较：

相同：都是用于检验多组样本是否来自同一总体

不同：ANOVA 的总体参数是平均数；K-W 检验的总体参数是中位数

K-W 检验的统计量与检验方法

设有 k 组样本，各自有  $n_1, n_2, \dots, n_k$  个统计量，且  $n = n_1 + n_2 + \dots + n_k$

1. 进行所有观察值的排序 (遇到相同值要平均)
2. 计算各组之等级和
3. 建立虚无与对立假设  $H_0: \eta_1 = \eta_2 = \dots = \eta_k$   $H_1: \eta_i (i = 1, 2 \dots k)$  不全相同
4. 统计量  $H = \frac{12}{n(n+1)} \sum_{i=1}^k \frac{r_i^2}{n_i} - 3(n+1)$
5. 检验方法与临界值  $H \sim \chi^2(k-1)$

注意：

- (1) 因为不须假设资料为常态分布，无母数分析常用于分析名目、有序的资料或资料分布未知的问题

- (2) K-W 检定与母数分析法中变异数分析的使用情况相似
- (3) K-W 检定各组样本数建议至少要 5 以上

#### KOLMOGOROV-SMIRNOV 检验

1. 1.K-S 检验以相对次数为依据，其检验效果不会因样本的大小而改变
2. 2.K-S 检验适用于小样本卡方检验 (**Chi-Squared Test**) 或无法进行的问题
3. 3.K-S 检验不必如卡方检验：假定理论次数在各细格中都必须  $\geq 5$ ，也不必将理论次数  $< 5$  的各组次数合并，能保有原资料的分布情形
4. 4.K-S 检验适用于顺序型资料，不适用于名义型资料。但是卡方检验适用于名义型资料。

K-S 检验步骤：

1. 1.建立假设：  
 $H_0$ : 此次分布与理论分布配适得当;  
 $H_1$ : 此次分布与理论分布配适不得当
2. 计算理论分布各阶段的累加概率  $F(x)$
3. 计算实际分布阶段的累加相对次数  $S(x)$
4. 计算各阶段  $|F(x) - S(x)|$
5. 找出 K-S 检验的检验统计量  $D = \max |F(x) - S(x)|$
6. 利用 K-S 检验表，有样本大小  $n$  及显著水平<sup>a</sup>，找出临界值  $D_{(\alpha/2, n)}$ ，当  $D > D_{(\alpha/2, n)}$  时，拒绝  $H_0$

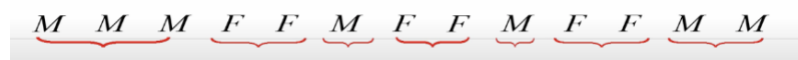
#### 第 31 章 无母数检定/非参检验三

##### 随机性检验(RUN TEST)

1. 观念：资料是否具有随机性，对于推论统计至关重要
2. 假设检验：  $H_0$ ：资料具有随机性；  $H_1$ ：资料不具有随机性
3. 又称连检验

检验方法：

1. 将资料分为两组。eg：男(M)，女(F)；不良品(G)，良品(B)。以中位数为界限，小雨中位数为 (-)，大于中位数为 (+)
2. 符号：  $n_1$  表示男生数量，  $n_2$  表示女生数量，  $v$  表示总共的串数（交错的串）



$$n_1 = 7 \quad n_2 = 6 \quad V = 6$$

小样本查表 ( $n_1, n_2 \leq 10$ )

大样本

1.  $n_1, n_2 > 10$ ，采用 Z 表

2.  $(\mu_V, \sigma_V^2)$
3.  $\mu_V = \frac{2n_1n_2}{n_1+n_2} + 1$       $\sigma_V^2 = \frac{2n_1n_2(2n_1n_2-n_1-n_2)}{(n_1+n_2)^2(n_1+n_2-1)}$
4. 转换为标准正态分布  $Z = \frac{V-\mu_V}{\sigma_V}$  检验

注意：

- (1) 统计量  $V = \text{runs (连) 的数}$
- (2) 如无假设为真，则来自两样本之分数序列应参杂排列，且有适当的 runs 数
- (3) 当至少个本之数目是大于 10 时，此 run 之抽样分配趋近常态，我们又可用 Z 分配表来做测定之基础

## FRIEDMAN 二因子检验

FRIEDMAN 二因子检验-----取代方差分析

1. 目的：比较 k 个相同总体是否具有相同的中位数
2. 方差分析要求总体为常态
3. FRIEDMAN 检验主要用于重复样本次序变数资料的非参数的统计方法。将个体重复接受 k 个实验条件后资料分成等级，便可以利用 FRIEDMAN 二因子等级方差分析进行分析。

检验步骤：

- (1) 建立虚无假设  $H_0: \eta_1 = \eta_2 = \dots = \eta_k$      $H_1: \eta_i (i = 1, 2 \dots k)$  不全相同
- (2) 将每一个集区观察值由小到大排序（个别排序，不要综合），再计算等级和  $R_i$
- (3) 统计量  $F_r = \frac{12}{nk(k+1)} \sum_{i=1}^n R_i^2 - 3n(k+1)$  服从自由度 (k-1) 的卡方分布

## 多个总体之中位数检验

1. 目的：与 Krustal Wallis 检验相同，都是比较 k 个总体是否具有相同的中位数
2. Krustal Wallis 检验要求样本来自 k 个独立、连续的资料；多个总体的中位数卡方检验没有限制
3. 以列联表，以中位数为分水岭，分成两组进行讨论

检验步骤：

- (1) 建立虚无假设  $H_0: \eta_1 = \eta_2 = \dots = \eta_k$      $H_1: \eta_i (i = 1, 2 \dots k)$  不全相同
- (2) 将 k 个总体中抽出的观察值排序，分成两组：a 大于中位数，b 小于中位数
- (3) 统计量  $\chi^2 = \sum_{i=1}^k \sum_{j=1}^2 \frac{o_{ij} - e_{ij}}{e_{ij}}$  服从自由度 (k-1) 的卡方分布

课程反思：

1. 多个总体中位数检验的想法
2. 若多个总体的中位数相同，则所有观察值混合排序，预期每个总体的观察值有一半低于混合排序的中位数，另一半高于混合排序的中位数。若不是如此卡方值将会变大，结论拒绝  $H_0$

## 交叉列联表 (contingency table)

- (1) 次数分布一次只能说明一个变数
- (2) 要说明两个或两个以上的变数使用交叉列联表

年龄	工作压力
----	------

	低	中	高
25 岁以下	20	18	22
25 岁 ~ 40 岁	50	46	44
40 岁 ~ 60 岁	58	63	59
60 岁以上	34	43	43

检定工作压力与年龄的关联性，采用卡方检定但并无适当的指标说明关联的强度（ $df = (r-1)(c-1)$ ）

#### PHI 系数

1.  $\phi$  系数是针对用来测量 2x2 的表格的关联强度
2.  $\phi$  系数是卡方统计值处以观察值个数后开平方根

$$\phi = \sqrt{\frac{\chi^2}{n}}$$

3. 若关联性低则值接近 0，当变量关联性高时  $\phi$  值未必在 1 内
4. 高度相关时所有的观察值会落在主对角线或次对角线上

#### 列联系数：contingency coefficient

评估任何大小表格的关系强度

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

列联系数范围：（0, 1）

列联系数适用于大于 2x2 的列联表

样本数越大，列联系数越低，Cramér's V 系数可以修正此问题

#### Cramér's V 系数 克莱姆系数

$$V = \sqrt{\frac{\chi^2}{n \cdot (k-1)}} \quad \text{其中 } k = \min(\text{列}, \text{行})$$

系数范围：（0, 1），靠近 1，则关联性强；反之靠近 0，关联性弱

#### 课程反思：

进行交叉列表分析时的步骤：

- (1) 利用卡方统计值来检定变数间有无关联性
- (2) 若不拒绝  $H_0$  虚无假设。则表示不具关联性
- (3) 若拒绝  $H_0$  表不具关联性，可以利用本单元介绍适当的统计值（ $\phi$  系数，列联系数 C 或 Cramer's 系数）来决定关系强度

## 第 32 章 统计应用

### 品质管制

品质：满足消费者需求的产品或服务

品质管控的目的：

商品品质的不稳定，会失去消费者的信赖，最后必然在市场遭受淘汰的命运

品质管制的方法：

- (1) SPC 统计的制程管理 (Statistical Process Control)

以管制图监控生产过程是否正确随时做出调整、继续的决策制造中的管制 (in- process control)

- (2) 允许抽样 (acceptance sampling)

抽取一批样本，检验品质以决定是否接受这批产品。制造后的管制 (after- process control)

品质变异的来源

1. 随机变异：

自然发生的变异，除非更换机器或原料否则无法消除（机器磨损、气温、湿度）

2. 非自然变异：

可通过调查改进（如：操作不当）

品质管制的重要性：

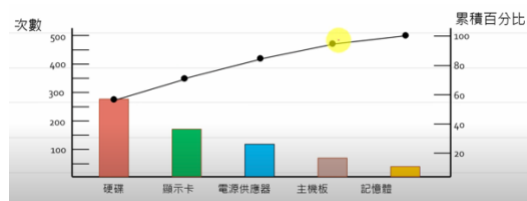
1. 提高产品竞争力，获取高收益
2. 降低成本，不良品会增加制造、维修、废料的处理成本
3. 不良品造成商誉损失，失去消费者的信赖，甚至被市场淘汰

### 诊断管制图

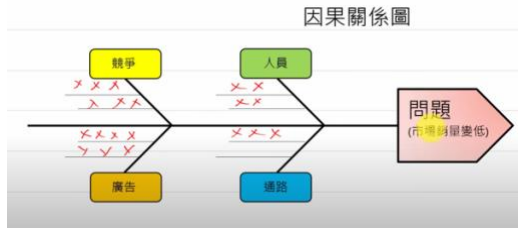
探索品质的常用方法：Pareto 分析图和鱼骨图

1. Pareto 分析图：记录产品缺失种类及数目的技术
2. 鱼骨图：强调问题因果可能的关系，对组织概念与问题解决有启发作用

Pareto 分析图



鱼骨图

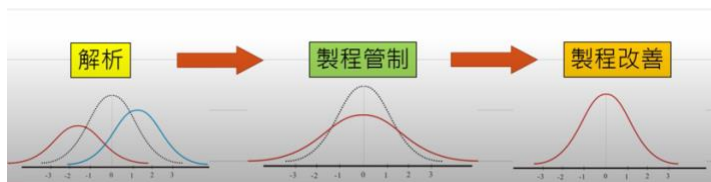


## 计量管制图

允许抽样 (acceptance sampling)

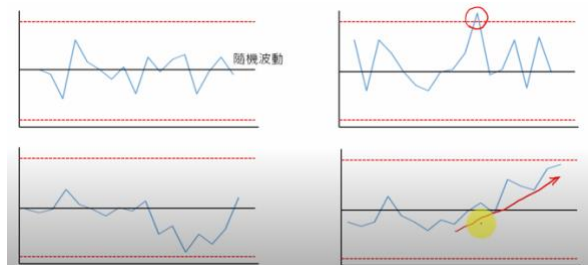
生产前后的整批检验，如果不合格数小于预先规定的数值则全批允收

统计制程管制 (Statistical Process Control)



管制图的目的：

- (1) 在监看制程是否处于稳定状态，利用管制图可以显示测定过程是否偏离统计控制的状况，并适时提出警讯
- (2) 管制上下决定出可接受变异范围



常用的管制图

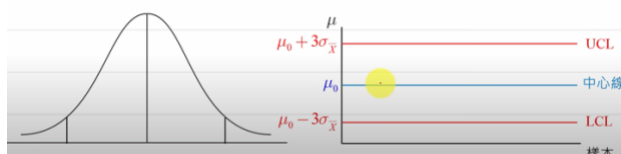
- (1) 平均值管制图：观察制程集中趋势（如： $\bar{x}$  管制图）
- (2) 全距管制图：观察制程分散趋势（如：R 管制图）

$\bar{x}$  管制图

定时抽取样本计算平均数 $\bar{x}$ ，品质管制人员利用统计的假设检验以了解产品制程是否在管制中

$H_0$ ：产品制程在管制中

$H_1$ ：产品制程不在管制中

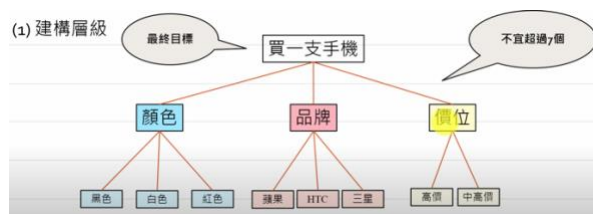


## AHP 阶层分析概率一

### AHP 介绍

- (1) 层级分析程序 (analytic hierarchy process)
- (2) 功能：使复杂问题分解成各个组成要素，再将这些要素依照关系分组成简明的层级结构系统
- (3) 目的：将复杂问题系统化，由不同层面给予层级分解，并透过量化的判断，协助决策者评估
- (4) 使用矩阵的特征向量 (Eigen vectors) 与特征值(Eigen values)的观念

#### (1) 建构层级



#### (2) 评估尺度 (问卷调查)

- (a) 每一个层级的上一层要素作为下一层级要素间的评估依据
  - (b) 分别评估该 2 个要素对评估准则的相对贡献度，问题分解为两两成对比较
- 若有多位受访者进行问卷采用几何平均数处理更合适

$$W = \sqrt[n]{w_1 \cdot w_2 \dots w_n}$$

#### (3) 建立成对比较矩阵

	價格	品牌	顏色
價格	1	3	5
品牌	$\frac{1}{3}$	1	2
顏色	$\frac{1}{5}$	$\frac{1}{2}$	1
	1.533	4.5	8

标准化成对矩阵:

	價格	品牌	顏色
價格	0.6522	0.6667	0.6250
品牌	0.2174	0.2222	0.2500
顏色	0.1304	0.1111	0.1250
	1.533	4.5	8

#### (4) 每列的平均称为各因素的权重



	價格	品牌	顏色
價格	0.6522	0.6667	0.6250
品牌	0.2174	0.2222	0.2500
顏色	0.1304	0.1111	0.1250

價格	0.64795
品牌	0.22987
顏色	0.12218

## 特征值与特征向量

$$A\vec{V} = \lambda\vec{V}$$

$\lambda$ 是特征值  $\vec{V}$ 是特征向量

$$A = \begin{bmatrix} 1 & 3 & 5 \\ 3 & 1 & 2 \\ 5 & 2 & 1 \end{bmatrix} \quad v = \begin{bmatrix} 0.64795 \\ 0.22987 \\ 0.12218 \end{bmatrix} \quad Av = \begin{bmatrix} 1 \cdot 0.64795 + 3 \cdot 0.22987 + 5 \cdot 0.12218 \\ 3 \cdot 0.64795 + 1 \cdot 0.22987 + 2 \cdot 0.12218 \\ 5 \cdot 0.64795 + 3 \cdot 0.22987 + 1 \cdot 0.12218 \end{bmatrix} = \begin{bmatrix} 1.94847 \\ 0.69021 \\ 3.66670 \end{bmatrix}$$
  

$$\begin{bmatrix} 1.94847 \\ 0.64795 \\ 0.69021 \\ 0.22987 \\ 0.36670 \\ 0.12218 \end{bmatrix} = \begin{bmatrix} 3.007130 \\ 3.002627 \\ 3.001318 \end{bmatrix} \quad \lambda = 3.003697 \quad C.I. = \frac{3.003697 - 3}{3 - 1} = 0.001848 \leq 0.1$$