

应用数理统计

胡奕公

武汉大学

目录

第一章 概率论复习	1
1.1 概率空间	1
1.2 随机变（向）量及其分布	3
1.2.1 随机变量及其分布	3
1.2.2 随机向量及其分布	5
1.2.3 边缘分布	7
1.2.4 条件分布	7
1.3 随机变量的独立性	8
1.4 随机变量函数的分布	8
1.4.1 单个随机变量函数的分布	8
1.4.2 随机向量函数的分布	9
1.5 随机变量的数字特征	11
1.5.1 数学期望	11
1.5.2 方差、协方差、相关系数	11
1.5.3 矩、协方差矩阵	13
1.5.4 两个重要不等式	13
1.6 特征函数	13
1.7 极限定理	15
1.7.1 大数定律	15
1.7.2 中心极限定理	16
第二章 数理统计的基本概念	19
2.1 总体与样本	19
2.2 统计量	20
2.3 三大统计抽样分布	21
2.4 正态总体抽样分布定理	22
第三章 参数估计	25
3.1 点估计	25
3.1.1 点估计的方法	25
3.1.2 估计量的评选标准	26
3.2 区间估计	27
3.2.1 置信区间	27
3.2.2 正态总体均值和方差的区间估计	28

第四章 假设检验	31
4.1 假设检验	31
4.2 正态总体参数的假设检验	31
4.3 似然比检验	33
4.4 非参数假设检验	33
4.4.1 分布拟合检验	33
4.4.2 独立性检验	34
第五章 回归分析	35
5.1 回归分析的基本概念	35
5.2 一元回归	35
5.2.1 一元线性回归	35
5.2.2 非线性回归	38
第六章 方差分析	41
6.1 单因素实验的方差分析	41
6.2 双因素方差分析	43
6.2.1 非重复实验双因素方差分析	43

第一章 概率论复习

1.1 概率空间

(1) **样本空间与事件域** 设 Ω 是样本空间, \mathcal{F} 是由 Ω 的一些子集构成的集类, 如果满足下列条件则 \mathcal{F} 是一个事件域, 每个时间发生的概率大小为 P , 三元体 (Ω, \mathcal{F}, P) 就构成一个概率空间。

- $\Omega \in \mathcal{F}$
- 若 $A \in \mathcal{F}$, 则 $\bar{A} \in \mathcal{F}$
- 若 $A_1, A_2, \dots, A_n, \dots \in \mathcal{F}$, 则

$$\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$$

(2) **概率的性质**

- $P(\phi) = 0$
- 可列可加性: 如果 $A_i \in \mathcal{F} (i = 1, 2, \dots, n)$ 且 $A_i \cap A_j = \phi (i \neq j)$, 则

$$P\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n P(A_i)$$

- 设 $A \in \mathcal{F}$, 则有 $P(A) = 1 - P(\bar{A})$
- 设 $A \in \mathcal{F}, B \in \mathcal{F}$, 如果 $A \supset B$, 则 $P(A - B) = P(A) - P(B), P(A) \geq P(B)$
- 设 $A \in \mathcal{F}, B \in \mathcal{F}$, 则

$$P(A \cup B) = P(A) + P(B) - P(AB) \leq P(A) + P(B)$$

可以推广到 n 个事件的情况

(3) **条件概率** 设 A, B 是两个随机事件, 且 $P(A) > 0$, 则称

$$P(B|A) = \frac{P(AB)}{P(A)} \quad (1.1.1)$$

为在事件 A 发生的条件下, 事件 B 发生的概率。

- 对于任意一个事件 B , $P(B|A) \geq 0$
- $P(\Omega|A) = 1$
- 设 B_1, B_2, \dots 互不相容, 则

$$P\left(\bigcup_{i=1}^{\infty} B_i \middle| A\right) = \sum_{i=1}^{\infty} P(B_i, A) \quad (1.1.2)$$

- $P(\bar{B}|A) = 1 - P(B|A)$

(i) **乘法公式** 设 $A, B \subset \Omega$, 当 $P(A) > 0$ 由

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

得到

$$P(AB) = P(B|A)P(A) = P(A|B)P(B) \quad (1.1.3)$$

推广得到

$$P(A_1, A_2, \dots, A_n) = P(A_1)P(A_2|A_1)P(A_3|A_2A_1) \cdots P(A_n|A_1A_2 \cdots A_{n-1}) \quad (1.1.4)$$

(ii) **全概率公式** 设随机试验 E 的样本空间为 Ω , A 为 E 的任意一事件, B_1, B_2, \dots, B_n 为 Ω 的一个划分, 且 $P(B_i) > 0$, 则

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \cdots + P(A|B_n)P(B_n) \quad (1.1.5)$$

在较复杂情况下直接计算 $P(A)$ 不易, 但 A 总是伴随着某些 B_i 出现, 适当地去构造这一组 B_i 往往可以使问题简化。

(iii) **贝叶斯公式** 设随机试验 E 的样本空间为 Ω , $A \subset \Omega$, B_1, B_2, \dots, B_n 为 Ω 的一个划分, $P(A) > 0, P(B_i) > 0$, 则

$$P(B_i|A) = \frac{P(AB_i)}{P(A)} = \frac{P(A|B_i)P(B_i)}{\sum_{i=1}^n P(A|B_i)P(B_i)} \quad (1.1.6)$$

它是在观察到事件 A 已发生的条件下, 寻找导致 A 发生的每个原因的概率。

(4) 事件的独立性

(i) **两个事件** 设 A, B 是两个事件, 如果如下等式成立

$$P(AB) = P(A)P(B) \quad (1.1.7)$$

则称事件 A, B 相互独立。

(ii) **三个事件** 对于 A, B, C 三个事件, 如果如下等式成立

$$\begin{aligned} P(AB) &= P(A)P(B) \\ P(AC) &= P(A)P(C) \\ P(BC) &= P(B)P(C) \end{aligned} \quad (1.1.8)$$

则称 A, B, C 两两独立; 如果满足

$$P(ABC) = P(A)P(B)P(C) \quad (1.1.9)$$

则称 A, B, C 互相独立。

(iii) **多个事件** 设 A_1, A_2, \dots, A_n 是 n 个事件, 若对任意的 $k (2 \leq k \leq n)$ 和任意一组 $1 \leq i_1 < i_2 < \cdots < i_k \leq n$ 都有

$$P(A_{i_1}, A_{i_2}, \dots, A_{i_k}) = P(A_{i_1})P(A_{i_2}) \cdots P(A_{i_k}) \quad (1.1.10)$$

成立, 则称 n 个事件 A_1, A_2, \dots, A_n 相互独立。

(iv) **可数无穷多个** 对于事件序列 $A_1, A_2, \dots, A_n, \dots$ 若他们之间任意有限个事件独立, 则称事件序列 $A_1, A_2, \dots, A_n, \dots$ 独立。

(v) **事件独立的性质** 若 A_1, A_2, \dots, A_n 独立, 则

- A'_1, A'_2, \dots, A'_n 独立, 其中 $A'_k = A_k \wedge \bar{A}_k$
- 将事件 A_1, A_2, \dots, A_n 分成 k 组, 设 B_1, B_2, \dots, B_n 分别由第 $1, 2, \dots, k$ 组内的 A_i 经过并、积、差、求余等运算所得, 则 B_1, B_2, \dots, B_n 独立。

1.2 随机变（向）量及其分布

1.2.1 随机变量及其分布

(1) **随机变量** 设随机试验的样本空间 $\Omega = \omega$, $\xi = \xi(\omega)$ 是定义在样本空间 Ω 上的**实值单值函数**, 称 $\xi = \xi(\omega)$ 是随机变量。

(i) **随机变量和普通函数的区别**

- 定义域不同: 随机变量定义在样本空间 Ω 上, 定义域可以是数也可以不是数; 而普通函数是定义在实数域上。
- 随机变量函数的取值在试验之前无法确定, 且取值有一定的概率; 而普通函数却没有。

(ii) **随机变量的分类**

- 离散型随机变量
- 连续型随机变量
- 混合型随机变量

(2) **分布函数** 设 X 是一个随机变量, x 是任意实数, 称函数

$$F(x) = P\{X < x\} \quad (-\infty < x < +\infty) \quad (1.2.1)$$

为 X 的分布函数。分布函数 $F(x)$ 的值就表示 X 落在区间 $(-\infty, x]$ 上的概率。

(i) **特点**

- 分布函数完整描述了随机变量的统计规律性
- 分布函数是一个普通实值函数

(ii) **性质** 以下三条性质是判断函数是否是分布函数的充要条件。

- 单调不减
- $0 \leq F(x) \leq 1$ 且

$$\begin{aligned} F(-\infty) &= \lim_{x \rightarrow -\infty} F(x) = 0 \\ F(+\infty) &= \lim_{x \rightarrow +\infty} F(x) = 1 \end{aligned} \quad (1.2.2)$$

- 右（左）连续性:

$$\begin{aligned} F(x-0) &= \lim_{x \rightarrow x-0} F(y) = F(x) \\ F(x+0) &= \lim_{x \rightarrow x+0} F(y) = F(x) \end{aligned} \quad (1.2.3)$$

(iii) **常用的概率公式**

- $P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a)$
- $P(X = x_0) = P(X \leq x_0) - P(X < x_0) = F(x_0) - F(x_0 - 0)$
- $P(X \geq x_0) = 1 - P(X < x_0) = 1 - F(x_0 - 0)$
- $P(X > x_0) = 1 - P(X \leq x_0) = 1 - F(x_0)$

1.2.1.1 离散型随机变量及其分布

(1) **离散型随机变量** 若随机变量 X 的全部可能取值是有限个或可列无限多个, 则称此随机变量是离散型随机变量。

(2) **分布律** 设离散型随机变量 X 的所有可能取值为 $x_k, k = (1, 2, \dots)$, X 取各个可能值的概率为 $P(X = x_k) = p_k$, p_k 满足下列条件, 则称 p_k 为离散型随机变量 X 的概率分布或分布律。

- $p_k \geq 0$
- $\sum_{k=1}^{\infty} p_k = 1$

(3) 常用的离散型随机变量分布(i) **0—1 分布** $P(X = k) = p^k(1-p)^{1-k}, k = 0, 1$ (ii) **二项分布** $P(X = k) = C_n^k p^k (1-p)^{n-k}, k = 0, 1, \dots, n$. 记为 $X \sim b(n, p)$

- 含义: n 重贝努里试验中出现“成功”次数 X 的概率分布。

- 退化为 0—1 分布: $n = 1$

(iii) **泊松分布** 分布律

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

其中 λ 是常数。记为 $X \sim P(\lambda)$ 。泊松分布在管理科学、运筹学以及自然科学的某些问题中都占有重要的地位。

- 排队问题: 在一段时间内窗口等待服务的顾客人数
- 生物存活的个数
- 放射的粒子数

1.2.1.2 连续型随机变量及其分布(1) **连续性随机变量** 如果随机变量 X 的分布函数为

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

其中被积函数 $f(t) \geq 0$ 则 X 为连续型随机变量, 称 $f(t)$ 为概率密度函数或概率密度。

(2) 概率密度的性质

- $f(x) \geq 0$
- $\int_{-\infty}^{\infty} f(x) dx = 1$
- $P(a < X \leq b) = F(b) - F(a) = \int_a^b f(x) dx$
- $P(X = a) = 0$
- 在 $f(x)$ 的连续点 x 处, 有

$$\begin{aligned} f(x) = F'(x) &= \lim_{\Delta x \rightarrow 0^+} \frac{F(x + \Delta x) - F(x)}{\Delta x} \\ &= \lim_{\Delta x \rightarrow 0^+} \frac{P(x < X \leq (x + \Delta x))}{\Delta x} \end{aligned} \quad (1.2.4)$$

- $P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b)$
- 若已知连续型随机变量 X 的密度函数为 $f(x)$, 则 X 在任意区间 G 上取值的概率为

$$P\{X \in G\} = \int_G f(x) dx \quad (1.2.5)$$

(3) 常见的连续型随机变量的分布(i) **均匀分布** X 的概率密度为:

$$f(x) = \begin{cases} \frac{1}{b-a}, & x \in (a, b) \\ 0, & x \in (-\infty, a] \cup [b, \infty) \end{cases} \quad (1.2.6)$$

则称 X 服从 (a, b) 上的均匀分布, 记为 $X \sim U(a, b)$ 。分布函数为

$$F(x) = \int_{-\infty}^x f(t) dt = \begin{cases} 0, & x < a \\ \frac{x-a}{b-a}, & a \leq x < b \\ 1, & x \geq b \end{cases} \quad (1.2.7)$$

(ii) 指数分布 X 的概率密度为:

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases} \quad (1.2.8)$$

则称 X 服从参数为 λ 的指数分布, 记为 $X \sim E(\lambda)$ 。分布函数为

$$F(x) = \begin{cases} 1 - e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases} \quad (1.2.9)$$

(iii) 正态分布 X 的概率密度为:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad -\infty < x < \infty \quad (1.2.10)$$

其中 μ 和 σ 为常数, 且 $\sigma > 0$, 则称 X 服从参数为 μ, σ 的正态分布或高斯分布, 记为 $X \sim N(\mu, \sigma^2)$ 。

(4) 正态分布

(i) 图形特点 正态分布的密度曲线是一条关于 μ 对称的钟形曲线, 特点是: “两头小, 中间大, 左右对称”

- μ 决定了图形的中心位置
- σ 决定了图形中峰的陡峭程度

(ii) 标准正态分布 $\mu = 0, \sigma = 1$ 的分布被称为标准正态分布, 其密度函数和分布函数常用 $\phi(x)$ 和 $\Phi(x)$ 表示

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

(iii) 定理 设 $X \sim N(\mu, \sigma^2)$, 则

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

1.2.2 随机向量及其分布

(1) n 维随机向量 设 (Ω, \mathcal{F}, P) 是一概率空间, $\xi_1(\omega), \xi_2(\omega), \dots, \xi_n(\omega)$ 是定义在这个概率空间上的 n 个随机变量, 称 $\xi(\omega) = (\xi_1(\omega), \xi_2(\omega), \dots, \xi_n(\omega))$ 为 n 维随机向量。

(2) n 维随机向量的联合分布 称 n 元函数

$$F_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n) = P(\xi_1 \leq x_1, \xi_2 \leq x_2, \dots, \xi_n \leq x_n)$$

为随机向量 $\xi = (\xi_1, \xi_2, \dots, \xi_n)$ 的联合分布函数。

(i) 二维随机变量分布函数的几何意义 将二维随机变量 (X, Y) 看平面上随机点的坐标, (X, Y) 落在区域 $\{x < x_2, y < y_2\}$ 中的概率为 $F_{X,Y}(x, y)$

(ii) 分布函数的性质

- 单调不减
- $0 < F < 1$
- 对于任意 x, y 有

$$F(x, -\infty) = \lim_{y \rightarrow -\infty} F(x, y) = 0 \quad (1.2.11)$$

$$F(-\infty, y) = \lim_{x \rightarrow -\infty} F(x, y) = 0 \quad (1.2.12)$$

- 右连续

1.2.2.1 离散型随机向量

(1) **离散型随机向量** 若随机向量 $\xi = (\xi_1, \xi_2, \dots, \xi_n)$ 只取有限个或可列个不同的向量值, 则称 ξ 为离散型随机向量。

(2) **分布列** 设 ξ 的所有可能取值为 $(x_{1i_1}, x_{2i_2}, \dots, x_{ni_n}), i_1, \dots, i_n = 1, 2, \dots$ 则称概率

$$p_{i_1, i_2, \dots, i_n} = P(\xi_1 = x_{1i_1}, \xi_2 = x_{2i_2}, \dots, \xi_n = x_{ni_n}) \quad (1.2.13)$$

为 ξ 的分布列。

(i) **分布列的性质**

- $p_{i_1, i_2, \dots, i_n} \geq 0$
- $\sum_{i_1, i_2, \dots, i_n} p_{i_1, i_2, \dots, i_n} = 1$

(ii) **分布函数**

$$F(x_1, x_2, \dots, x_n) = \sum_{x_{1i_1} < x_1, x_{2i_2} < x_2, \dots, x_{ni_n} < x_n} p_{i_1, i_2, \dots, i_n} \quad (1.2.14)$$

1.2.2.2 连续型随机向量

(1) **连续型随机向量** 若存在非负函数 $p(x_1, x_2, \dots, x_n)$ 使得分布函数可表示为

$$F(x_1, x_2, \dots, x_n) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \dots \int_{-\infty}^{x_n} f(s_1, s_2, \dots, s_n) ds_1 ds_2 \dots ds_n$$

则称 ξ 为连续型随机变量。函数 $f(s_1, s_2, \dots, s_n)$ 称为 ξ 的分布密度或密度函数, 满足条件下列条件。反之, 若有满足这两条性质的 n 元函数 $f(s_1, s_2, \dots, s_n)$ 则他一定是某一个 n 维随机变量的密度函数。

- $f(s_1, s_2, \dots, s_n) > 0$
- $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f(s_1, s_2, \dots, s_n) ds_1 ds_2 \dots ds_n = 1$

(2) **概率** 对于 R^n 上某一区域 B 有

$$P\{(\xi_1, \xi_2, \dots, \xi_n) \in B\} = \iint \dots \int_B f(s_1, s_2, \dots, s_n) dx_1, dx_2, \dots, dx_n \quad (1.2.15)$$

(3) **多元正态分布** 设 $\mathbf{B} = (b_{ij})$ 是 n 阶正定对称矩阵, $\mathbf{B}^{-1} = (r_{ij})$ 是他的逆矩阵, $|\mathbf{B}|$ 表示 \mathbf{B} 的行列式, $\mathbf{a} = (a_1, a_2, \dots, a_n)^T$ 是一个实值列向量, 以函数

$$f(x_1, x_2, \dots, x_n) = \frac{1}{(2\pi)^{\frac{n}{2}} |\mathbf{B}|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2} \sum_{i,j=1}^n r_{ij} (x_i - \mathbf{a}_i)(x_j - \mathbf{a}_j) \right\} \quad (1.2.16)$$

为密度函数的概率分布, 称为 n 元正态分布, 简记为 $N(\mathbf{a}, \mathbf{B})$ 。向量形式为

$$f(x_1, x_2, \dots, x_n) = \frac{1}{(2\pi)^{\frac{n}{2}} |\mathbf{B}|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \mathbf{a})^T \mathbf{B}^{-1} (\mathbf{x} - \mathbf{a}) \right\} \quad (1.2.17)$$

其中, $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$ 。

(i) **二元正态分布** 当 $n = 2$ 时, (X, Y) 的概率密度为

$$f(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp \left\{ -\frac{1}{2(1-\rho^2)} \cdot \left[\frac{(x-\mu_1)^2}{\sigma_1^2} - 2\rho \frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2} \right] \right\} \quad (1.2.18)$$

其中, $\mu_1, \mu_2, \sigma_1, \sigma_2, \rho$ 都是常数, 且 $\sigma_1 > 0, \sigma_2 > 0, -1 < \rho < 1$, 则称 (X, Y) 服从参数为 $\mu_1, \mu_2, \sigma_1, \sigma_2, \rho$ 的二维正态分布, 记为

$$(X, Y) \sim N(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$$

1.2.3 边缘分布

(1) **边缘分布** 设 (X, Y) 是二维随机变量, X 的分布函数 $F_X(x)$ 称为 (X, Y) 关于 X 的边缘分布函数。

$$F_X(x) = P\{X \leq x\} = P\{X \leq x, Y < \infty\} = F(x, \infty)$$

同理

$$F_Y(y) = F(\infty, y)$$

(2) **边缘分布率**

$$P\{X = x_i\} = \sum_{j=1}^{\infty} P\{X = x_i, Y = y_j\} \quad (1.2.19)$$

$$P\{Y = y_i\} = \sum_{j=1}^{\infty} P\{X = x_j, Y = y_i\} \quad (1.2.20)$$

(3) **边缘概率密度**

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy \quad (1.2.21)$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx \quad (1.2.22)$$

(4) **二元正态分布的性质**

- 二维正态分布的两个边际分布都是一维正态分布
- 两个边缘分布中的参数与二元正态分布中的常数 ρ 无关
- 如果 $(X_1, Y_1) \sim N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho_1)$, $(X_2, Y_2) \sim N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho_2)$, 则 (X_1, Y_1) 与 (X_2, Y_2) 的分布不相同, 但是 X_1 与 X_2 分布相同, Y_1 与 Y_2 分布相同。
- 一般来讲不能由边缘分布求联合分布

1.2.4 条件分布

(1) **离散型** 设 (ξ, η) 是离散型随机向量, 他的联合分布为

$$p_{ij} = P(\xi = x_i, \eta = y_j)$$

若已知 $P(\xi = x_i) > 0$, 则在 $\xi = x_i$ 下, 事件 $\eta = y_j$ 的条件概率为

$$P\{\eta = y_j | \xi = x_i\} = \frac{P\{\xi = x_i, \eta = y_j\}}{P\{\xi = x_i\}} = \frac{p_{ij}}{\sum_j p_{ij}}$$

称为 η 关于 ξ 的条件分布。

(2) **连续型** 设 (ξ, η) 是连续随机向量, 他的联合密度为 $f(x, y)$ 若已知 $p_1(x) > 0$ 则称

$$F(y|x) = P\{\eta < y | \xi = x\} = \frac{\int_{-\infty}^y f(x, v) dv}{f_1(x)} = \int_{-\infty}^y \frac{f(x, v)}{f_1(x)} dv$$

称为 η 关于 ξ 的条件分布。

- (3) **二元正态分布** 设随机向量 (ξ, η) 服从正态分布, 则在 $\xi = x$ 的条件下, η 的条件分布也是正态分布, 其中第一个参数是 x 的线性函数, 第二个参数与 x 无关。

$$N\left(b + r \frac{\sigma_2}{\sigma_1}(x - a), \sigma_2^2(1 - r^2)\right)$$

1.3 随机变量的独立性

- (1) **二维随机变量的独立性** 若二维随机变量 (X, Y) 对任意实数 x, y 都有

$$P\{X \leq x, Y \leq y\} = P\{X \leq x\}P\{Y \leq y\}$$

即

$$F(x, y) = F_X(x)F_Y(y)$$

成立, 则称随机变量 X 与 Y 是相互独立的。

- 当 X, Y 独立时, 由 X, Y 的边缘分布可以唯一决定 (X, Y) 的联合分布。可直接推广至两个以上随机变量的相互独立性。
- 二元正态随机变量边缘分布相互独立的充分必要条件为

$$\rho = 0$$

- (2) **n 维随机变量的独立性** 设 X_1, X_2, \dots, X_n 是定义在概率空间 Ω, \mathcal{F}, P 上的 n 个随机变量, 若对于任意实数 x_1, x_2, \dots, x_n , 有

$$P\{X_1 < x_1, X_2 < x_2, \dots, X_n < x_n\} = P\{X_1 < x_1\}P\{X_2 < x_2\} \cdots P\{X_n < x_n\}$$

则称 X_1, X_2, \dots, X_n 相互独立。联合分布函数

$$F(x_1, x_2, \dots, x_n) = F_1(x_1)F_2(x_2) \cdots F_n(x_n)$$

(i) **充要条件**

- 离散型:

$$P\{X_1 = x_1, X_2 = x_2, \dots, X_n = x_n\} = P\{X_1 = x_1\}P\{X_2 = x_2\} \cdots P\{X_n = x_n\}$$

- 连续型:

$$f(x_1, x_2, \dots, x_n) = f_1(x_1)f_2(x_2) \cdots f_n(x_n)$$

(ii) **性质** 若随机变量 X_1, X_2, \dots, X_n 相互独立, 则

- 其中任意 $m (2 \leq m \leq n)$ 个随机变量也独立
- 他们的函数 $g_i(X_i)$ 也是随机变量, 也相互独立
- 函数 $g_i(X_{j_1}^{(i)}, X_{j_2}^{(i)}, \dots, X_{j_{t_i}}^{(i)})$ 是随机变量, 记 $A_i = \{j_1^{(i)}, j_2^{(i)}, \dots, j_{t_i}^{(i)}\}$, 若对任意 i, j 有 $A_i A_j = \Phi$ 则 $g_i(X_{j_1}^{(i)}, X_{j_2}^{(i)}, \dots, X_{j_{t_i}}^{(i)})$ 也是相互独立的。这类函数是非常宽泛的一类函数。

1.4 随机变量函数的分布

1.4.1 单个随机变量函数的分布

- (1) **离散型** 设 X 是离散型随机变量, 且概率函数为

$$P\{X = x_i\} = p_i, \quad i = 1, 2, \dots$$

则 $Y = f(X)$ 是离散型随机变量, 其分布律为

$$P\{Y = f(x_i)\} = p_i, \quad i = 1, 2, \dots$$

其中 $f(x_i)$ 相同的值的概率应相加。

(2) 连续型 设 X 为连续型随机变量, 其概率密度函数为 $f(x)$

(i) 反函数法

- 若 $y = g(x)$ 严格单调, 反函数 $g^{-1}(y)$ 有连续导数, 则 $Y = g(X)$ 是具有如下分布密度的连续型随机变量

$$f_Y(y) = \begin{cases} f(g^{-1}(y)) |(g^{-1}(y))'|, & \alpha < y < \beta \\ 0, & y \leq \alpha, y \geq \beta \end{cases}$$

其中

$$\alpha = \min \{g(-\infty), g(+\infty)\}$$

$$\beta = \max \{g(-\infty), g(+\infty)\}$$

- 若 $y = g(x)$ 在 I_1, I_2, \dots 不相重叠的区间上逐段严格单调, 其反函数分别为 $h_1(y), h_2(y), \dots$, 他们在各自区间上有连续导数, 则 $Y = g(X)$ 是连续型随机变量, 其分布密度在相应区间内为

$$f(h_1(y)) |h'_1(y)| + f(h_2(y)) |h'_2(y)| + \dots$$

- 一般地, 有

$$f_Y(y) = \frac{f_X(x_1)}{\left|\frac{dy}{dx}\right|_{x=x_1}} + \frac{f_X(x_2)}{\left|\frac{dy}{dx}\right|_{x=x_2}} + \dots + \frac{f_X(x_n)}{\left|\frac{dy}{dx}\right|_{x=x_n}}$$

(ii) 分布函数微分法 先用定义求 Y 的分布函数 $F_Y(y)$ 再求导得到密度函数 $f_Y(y)$

$$F_Y(y) = P\{Y \leq y\} = P\{g(X) \leq y\} = \int_S f_X(x) dx$$

其中 $S = \{x | g(x) \leq y\}$ 而 $Y = g(X)$ 的分布密度为

$$f_Y(y) = \frac{dF_Y(y)}{dy}$$

(3) 标准化随机变量 若随机变量 X 有有限方差 $DX > 0$, 则

$$Y = \frac{X - EX}{\sqrt{DX}}$$

满足 $EY = 0, DY = 1$ 称为 X 的标准化随机变量。

1.4.2 随机向量函数的分布

(1) 离散型 设 (X, Y) 是离散型随机变量, 且概率函数为

$$P\{X = x_i, Y = y_j\} = p_{ij}, \quad i, j = 1, 2, \dots$$

则 $Z = g(X, Y)$ 是离散型随机变量, 其分布律为

$$P\{Z = g(x_i, y_j)\} = p_{ij}, \quad i, j = 1, 2, \dots$$

其中 $g(x_i, y_j)$ 相同的值的概率应相加。

(2) 连续型

- (i) **分布函数微分法** 设 (X, Y) 是二位连续性随机变量, 其联合概率密度为 $f(x, y)$, 则 $Z = g(X, Y)$ 的分布函数为

$$F_Z(z) = P\{Z < z\} = P\{g(x, y) < z\} = \iint_D f(x, y) dx dy = \iint_{g(x, y) < z} f(x, y) dx dy$$

- **和的分布**: 设 $Z = X + Y$ 则 Z 的密度函数为

$$f_Z(z) = \int_{-\infty}^{+\infty} f(z - y, y) dy = \int_{-\infty}^{+\infty} f(x, z - x) dx \quad (1.4.1)$$

当 X, Y 独立时, $f_Z(z) = f_X(x) * f_Y(y)$ 。

- **最大最小值的分布**: 设 $M = \max(X_1, X_2, \dots, X_n)$, $N = \min(X_1, X_2, \dots, X_n)$, 则 M, N 的分布函数分别为

$$F_M(z) = F_{X_1}(z) F_{X_2}(z) \cdots F_{X_n}(z) \quad (1.4.2)$$

$$F_N(z) = 1 - [1 - F_{X_1}(z)][1 - F_{X_2}(z)] \cdots [1 - F_{X_n}(z)] \quad (1.4.3)$$

当 X_1, X_2, \dots, X_n 独立同分布时,

$$F_M(z) = [F_X(z)]^n \quad (1.4.4)$$

$$F_N(z) = 1 - [1 - F_X(z)]^n \quad (1.4.5)$$

(ii) 积分转换法

- **二元函数积分转换法**: 设 X_1, X_2 的联合密度函数为 $f(x_1, x_2)$, 若对于函数

$$\begin{cases} y_1 = g_1(X_1, X_2) \\ y_2 = g_2(X_1, X_2) \end{cases}$$

如果对任何有界连续的二元函数 $h = g(y_1, y_2)$ 满足以下条件

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(g_1(x_1, x_2), g_2(x_1, x_2)) f(x_1, x_2) dx_1 dx_2 = \iint_D h(y_1, y_2) p(y_1, y_2) dy_1 dy_2 \quad (1.4.6)$$

则 (Y_1, Y_2) 的联合密度函数为

$$f_{Y_1, Y_2}(y_1, y_2) = \begin{cases} p(y_1, y_2), & (y_1, y_2) \in D \\ 0, & (y_1, y_2) \notin D \end{cases} \quad (1.4.7)$$

- **多元函数积分转换法**: 设 X_1, X_2, \dots, X_n 的联合密度函数为 $f(x_1, x_2, \dots, x_n)$, 函数

$$y = g(x_1, x_2, \dots, x_n)$$

如果对任意有界连续函数 $h(y)$ 满足以下条件

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(g(x_1, x_2, \dots, x_n)) f(x_1, x_2, \dots, x_n) dx_1 dx_2 \cdots dx_n = \int_D h(y) p(y) dy \quad (1.4.8)$$

则 $Y = g(X_1, X_2, \dots, X_n)$ 的密度函数为

$$f_Y(y) = \begin{cases} p(y), & y \in D \\ 0, & y \notin D \end{cases} \quad (1.4.9)$$

1.5 随机变量的数字特征

1.5.1 数学期望

(1) 随机变量的数学期望 设随机变量 X 的分布函数为 $F_X(x)$ ，若积分

$$\int_{-\infty}^{+\infty} x dF(x)$$

绝对收敛，则该积分值为随机变量 X 的期望，记为 $E(X)$ 。

- 离散型

$$E(X) = \sum_{k=1}^{\infty} x_k p(x_k)$$

- 连续型

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx$$

(2) 随机变量函数的数学期望 设随机变量 Y 是随机变量 X 的函数 $Y = g(X)$ ，随机变量 X 的分布函数为 $F_X(x)$ ，则

$$E(Y) = \int_{-\infty}^{\infty} g(x) dF(x)$$

- 离散型

$$E(Y) = \sum_{k=1}^{\infty} g(x_k) p(x_k)$$

- 连续型

$$E(Y) = \int_{-\infty}^{\infty} g(x) f(x) dx$$

(3) 二维随机向量函数的数学期望 设 (X, Y) 为二维随机向量，分布函数为 $F_{X,Y}(x, y)$ ，又函数 $g(x, y)$ 在 R^2 上连续，则 $Z = g(X, Y)$ 的数学期望为

$$E(Z) = E[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) dF(x, y)$$

要求广义二重积分是绝对收敛的。

1.5.2 方差、协方差、相关系数

(1) 方差 设 X 是一个随机变量，若 $E((X - E(X))^2)$ 存在，则称

$$D(X) = E((X - E(X))^2)$$

是 X 的方差，方差的算术平方根 $\sqrt{D(X)} = \sigma(X)$ 称为标准差。

(i) 方差的含义 方差刻画了随机变量的取值对于其数学期望的离散程度。

(ii) 方差的运算率

- $D(X) = E(X^2) - E^2(X)$
- $D(aX + b) = a^2 D(X)$
- $D(X \pm Y) = D(X) + D(Y) \pm 2 \text{Cov}(X, Y)$
- 设 $X = (X_1, X_2, \dots, X_n)^T$ 为 n 元随机向量， $E(X) = \mathbf{a}$ ， $D(X) = \mathbf{B}$ ，
 - 对于 $Y = \mathbf{l}^T X$ ， $\mathbf{l} = (l_1, l_2, \dots, l_n)$ ，有 $E(Y) = \mathbf{l}^T \mathbf{a}$ ， $D(Y) = \mathbf{l}^T \mathbf{B} \mathbf{l}$
 - 对于 $Y = \mathbf{C}^T X$ ， $\mathbf{C}_{m \times n} = (c_{ij})_{m \times n}$ ，有 $E(Y) = \mathbf{C} \mathbf{a}$ ， $D(Y) = \mathbf{C} \mathbf{B} \mathbf{C}^T$

表 1.1 常见分布的期望和方差

分布	概率质量（密度）函数	期望	方差
0-1 分布	$1(p)$	p	$p(1-p)$
二项分布	$b(n, p)$	p	$np(1-p)$
泊松分布	$\pi(\lambda)$	λ	λ
均匀分布	$U[a, b]$	$\frac{b-a}{2}$	$\frac{(b-a)^2}{12}$
指数分布	$e(\theta)$	θ	θ^2
正态分布	$N(\mu, \theta^2)$	μ	θ^2

(2) **协方差** 设二维随机变量 (X, Y) 若 $E((X - E(X))(Y - E(Y)))$ 存在, 则称他为 X 与 y 的协方差, 记为 $\text{Cov}(X, Y)$ 。

(i) **协方差的性质**

- $\text{Cov}(X, X) = D(X)$
- $\text{Cov}(X, Y) = \text{Cov}(Y, X)$
- $\text{Cov}(a_1X + b_1, a_2Y + b_2) = a_1a_2\text{Cov}(X, Y)$
- $\text{Cov}(X_1 + X_2, Y) = \text{Cov}(X_1, Y) + \text{Cov}(X_2, Y)$
- $\text{Cov}(X, Y) = E(X, Y) - E(X)E(Y)$
- $\text{Cov}(a, Y) = 0$

(ii) **协方差的含义** 协方差的大小在一定程度上反映了 X 和 Y 相互间的关系, 但它还受 X 与 Y 本身度量单位的影响。

(3) **相关系数** 设 $D(X) > 0, D(Y) > 0$ 则称

$$\rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sqrt{D(X)}\sqrt{D(Y)}} \quad (1.5.1)$$

为随机变量 X 和 Y 的相关系数。

(i) **性质**

- $\rho_{X,Y} \leq 1$
- $\rho_{X,Y} = 1$ 的充分必要条件是 X 和 Y 以概率 1 呈线性关系。

(ii) **含义** 相关系数是 X 和 Y 间线性关系的一种度量

- $|\rho_{X,Y}| \rightarrow 1$, X 和 Y 间的线性关系越显著
- $|\rho_{X,Y}| \rightarrow 0$, X 和 Y 间的线性关系越不显著

(iii) **四个等价命题**

- $\rho_{X,Y} = 0 \Leftrightarrow X$ 和 Y 不相关
- $E(XY) = E(X)E(Y)$
- $\text{Cov}(X, Y) = 0$
- $D(X \pm Y) = D(X) + D(Y)$

(iv) **不相关性与独立性的关系**

- 不相关: X 和 Y 之间没有线性关系, 并不代表没有其它关系;
- 独立: X 和 Y 之前完全没有任何关系

- 独立 \Rightarrow 不相关, 但一般情况下不相关 \nRightarrow 独立
- 若 (X, Y) 服从二元正太分布, 则 X 和 Y 相关 $\Leftrightarrow X$ 和 Y 独立

1.5.3 矩、协方差矩阵

(1) 矩 设 X 和 Y 是随机变量, 有 $k, l = 1, 2, \dots$, 则

- (i) (原点) 矩 若 $E(X^k)$ 存在, 则称它为 X 的 k 阶 (原点) 矩;
- (ii) 中心矩 若 $E((X - E(X))^k)$ 存在, 则称它为 X 的 k 阶中心矩;
- (iii) 混合 (原点) 矩 若 $E(X^k Y^l)$ 存在, 则称它为 X 和 Y 的 $k + l$ 阶混合 (原点) 矩;
- (iv) 混合 (原点) 矩 若 $E((X - E(X))^k (Y - E(Y))^l)$ 存在, 则称它为 X 和 Y 的 $k + l$ 阶混合中心矩;

(2) 协方差矩阵 若 n 维随机变量的 n^2 个二阶中心矩都存在, 将他们排列成矩阵形式, 称为 (X_1, X_2, \dots, X_n) 的协方差矩阵

$$\mathbf{B} = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{nn} \end{bmatrix} \quad (1.5.2)$$

(i) 性质

- 对称矩阵
- 非负定矩阵。对于任一 n 元实列向量 \mathbf{t} 有

$$\mathbf{t}^T \mathbf{B} \mathbf{t} \geq 0$$

1.5.4 两个重要不等式

(1) 切比雪夫不等式 对任意具有有限方差的随机变量 X , 都有对 $\forall \epsilon > 0$ 不等式

$$P\{|X - E(X)| \geq \epsilon\} \geq \frac{D(X)}{\epsilon^2} \quad (1.5.3)$$

或

$$P\{|X - E(X)| < \epsilon\} \geq 1 - \frac{D(X)}{\epsilon^2} \quad (1.5.4)$$

(2) A.L.Cauchy - Schwarz 不等式 设随机向量 (X, Y) 满足 $E(X^2) < \infty, E(Y^2) < \infty$ 则有

$$[E(XY)]^2 \leq E(X^2)E(Y^2) \quad (1.5.5)$$

推论:

$$\text{Cov}^2(X, Y) = [E(X - E(X))(Y - E(Y))]^2 \leq E((X - E(X))^2)E((Y - E(Y))^2) = D(X)D(Y) \quad (1.5.6)$$

1.6 特征函数

(1) 复随机变量 设 X, Y 是概率空间 $\{\Omega, \mathcal{F}, P\}$ 上的实随机变量, 则 $Z = X + iY$ 称为复随机变量。

- (i) 独立性 如果 $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ 互相独立, 就称复随机变量 $X_1 + iY_1, X_2 + iY_2, \dots, X_n + iY_n$ 互相独立。
- (ii) 期望 $E(Z) = E(X) + iE(Y)$

(2) **特征函数** 设 X 的分布函数为 $F_X(x)$ 称

$$\phi(t) = E(e^{itX}) = \int_{-\infty}^{\infty} e^{itx} dF(x) = \begin{cases} \sum_{j=1}^{\infty} e^{itx_j} p_j \\ \int_{-\infty}^{\infty} e^{itx} f(x) dx \end{cases} \quad (1.6.1)$$

为 X 的特征函数。

(i) **存在性** 因为 $e^{itx} = \cos tx + i \sin tx$ 即 $|e^{itx}| = 1$ 则 $E(e^{itx})$ 总存在。

(ii) **性质**

- $\phi(0) = 1, |\phi(t)| \leq \phi(0), \phi(-t) = \overline{\phi(t)}$
- 特征函数 $\phi(t)$ 在 \mathbb{R} 上一致连续
- 特征函数 $\phi(t)$ 非负定, 对 $\forall t_1, t_2, \dots, t_n \in \mathbb{R}$ 及 $\forall a_1, a_2, \dots, a_n \in \mathbb{C}$

$$\sum_{k=1, j=1}^n \phi(t_k - t_j) a_k \overline{a_j} \geq 0 \quad (1.6.2)$$

- 设 a, b 是常数, $Y = aX + b$ 则

$$\phi_Y(t) = e^{itb} \phi_X(at) \quad (1.6.3)$$

- 随机变量 X, Y 独立, 则

$$\phi_{X+Y}(t) = \phi_X(t) \phi_Y(t) \quad (1.6.4)$$

可以推广至多个: 随机变量 X_1, X_2, \dots, X_n 相互独立, 则

$$\phi_{\sum_{i=1}^n X_i}(t) = \prod_{i=1}^n \phi_{X_i}(t) \quad (1.6.5)$$

- 设随机变量的 n 阶原点矩存在, 则他的特征函数可以微分 n 次, 且

$$E(X^k) = (-i)^k \phi_X^{(k)}(0) \quad (1.6.6)$$

$$\phi^{(k)}(0) = i^k E(X^k) \quad (1.6.7)$$

其中

$$\phi^{(k)}(t) = \frac{d^k}{dt^k} \int_{-\infty}^{\infty} e^{itx} dF(x) \quad (1.6.8)$$

- 特征函数的凸组合式特征函数。设 $\{\phi_n\}$ 是特征函数, 若 $\lambda_n \geq 0, \sum \lambda_n = 1$ 则称 $\sum \lambda_n \phi_n$ 也是特征函数。
- 特征函数的乘积也是特征函数。 $\prod \phi_n$ 是独立和 $\sum X_n$ 的特征函数。
- 唯一性定理: 分布函数由其特征函数唯一确定。
- 逆转公式: 若特征函数 $\phi(t)$ 绝对可积且相应的分布函数 $F(x)$ 可导且导函数连续, 则有

$$F'(x) = f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \phi(t) dt \quad (1.6.9)$$

对于任意分布函数, 有

$$F(x_2) - F(x_1) = \lim_{T \rightarrow \infty} \frac{1}{2\pi} \int_{-T}^T \frac{e^{-itx_1} - e^{-itx_2}}{it} \phi(t) dt \quad (1.6.10)$$

(3) **多元特征函数** 若随机向量 $(\xi_1, \xi_2, \dots, \xi_n)$ 的分布函数为 $F(x_1, x_2, \dots, x_n)$, 特征函数为

$$\phi(t_1, t_2, \dots, t_n) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} e^{i(t_1 x_1 + t_2 x_2 + \dots + t_n x_n)} dF(x_1, x_2, \dots, x_n) = E(e^{it^T X}) \quad (1.6.11)$$

(i) 性质

- 在 \mathbb{R}^n 中一致连续, 且

$$|\phi(t_1, t_2, \dots, t_n)| \leq \phi(0, 0, \dots, 0) = 1 \quad (1.6.12)$$

$$\phi(-t_1, -t_2, \dots, -t_n) = \overline{\phi(t_1, t_2, \dots, t_n)} \quad (1.6.13)$$

- 如果 $\phi(t_1, t_2, \dots, t_n)$ 是 $(\xi_1, \xi_2, \dots, \xi_n)$ 的特征函数, 则 $\eta = a_1\xi_1 + a_2\xi_2 + \dots + a_n\xi_n$ 的特征函数为

$$\phi_\eta(t) = \phi_\xi(a_1\mathbf{t}, a_2\mathbf{t}, \dots, a_n\mathbf{t}) = \phi_\xi(\mathbf{t}\mathbf{a}) \quad (1.6.14)$$

- 如果 $E(\xi_1^{k_1}\xi_2^{k_2}\dots\xi_n^{k_n})$ 存在, 则

$$E(\xi_1^{k_1}\xi_2^{k_2}\dots\xi_n^{k_n}) = i^{-\sum_{j=1}^n k_j} \left[\frac{\partial^{k_1+k_2+\dots+k_n} \phi(t_1, t_2, \dots, t_n)}{\partial t_1^{k_1} \partial t_2^{k_2} \dots \partial t_n^{k_n}} \right]_{t_1=t_2=\dots=t_n=0} \quad (1.6.15)$$

- 如果 $\phi(t_1, t_2, \dots, t_n)$ 是 $(\xi_1, \xi_2, \dots, \xi_n)$ 的特征函数, 则 $k(k < n)$ 维随机变量的特征函数是

$$\phi_{1,2,\dots,k} \leq \phi(t_1, t_2, \dots, t_n, 0, \dots, 0) \quad (1.6.16)$$

- 逆转公式: 如果 $\phi(t_1, t_2, \dots, t_n)$ 是 $(\xi_1, \xi_2, \dots, \xi_n)$ 的特征函数, 而 $F(x_1, x_2, \dots, x_n)$ 是它的分布函数, 则

$$P\{a_k < \xi_k < b_k\} = \lim_{T_j \rightarrow \infty} \frac{1}{(2\pi)^n} \int_{-T_1}^{T_1} \int_{-T_2}^{T_2} \dots \int_{-T_n}^{T_n} \prod_{k=1}^n \frac{e^{-it_k a_k} - e^{-it_k b_k}}{it_k} \cdot \phi(t_1, t_2, \dots, t_n) dt_1, dt_2, \dots, dt_n \quad (1.6.17)$$

其中, $k = 1, 2, \dots, n$, a_k 和 b_k 都是任意实数, 但满足 $(\xi_1, \xi_2, \dots, \xi_n)$ 落在平行体 $a_k \leq \xi_k \leq b_k, k = 1, 2, \dots, n$ 的面上的概率为零。

- 如果 $\phi(t_1, t_2, \dots, t_n)$ 是 $(\xi_1, \xi_2, \dots, \xi_n)$ 的特征函数, 而 ξ_j 的特征函数为 $\phi_j(t)$ 则随机变量 $\xi_1, \xi_2, \dots, \xi_n$ 相互独立的重要条件为

$$\phi(t_1, t_2, \dots, t_n) = \phi_{\xi_1}(t_1)\phi_{\xi_2}(t_2)\dots\phi_{\xi_n}(t_n) \quad (1.6.18)$$

- 设随机向量 $\xi = (\xi_1, \xi_2, \dots, \xi_n)$, $\eta = (\eta_1, \eta_2, \dots, \eta_m)$ 的特征函数分别为 $\phi_1(t_1, t_2, \dots, t_n)$, $\phi_2(u_1, u_2, \dots, u_m)$, 则 ξ 与 η 独立的充分必要条件是

$$\phi(t_1, t_2, \dots, t_n, u_1, u_2, \dots, u_m) = \phi_1(t_1, t_2, \dots, t_n)\phi_2(u_1, u_2, \dots, u_m) \quad (1.6.19)$$

- 连续性定理: 若特征函数列 $\{f_k(t_1, t_2, \dots, t_n)\}$ 收敛于某个连续函数 $f(t_1, t_2, \dots, t_n)$, 则函数 $f(t_1, t_2, \dots, t_n)$ 一定为某个分布函数所对应的特征函数。
- 随机向量 $X = (X_1, X_2, \dots, X_n)$ 的特征函数为 $\phi_X(t)$, 则随机变量 $Y = \mathbf{A}X + \mathbf{B}$ 的特征函数是

$$\phi_Y(t) = \phi_X(\mathbf{A}^T \mathbf{t}) e^{i\mathbf{t}\mathbf{B}} \quad (1.6.20)$$

1.7 极限定理

1.7.1 大数定律

以 μ_n 记实验中 A 出现的次数, 当 $n \rightarrow \infty$ 时, 频率 $\frac{1}{n}\mu_n$ 趋近于常数 $P(A)$ 。已知

$$E(\mu_n/n) = P(A) = p$$

$$D(\mu_n/n) = \frac{pq}{n}$$

标准化变量

$$\xi_n = \frac{\mu_n - np}{\sqrt{npq}}$$

则

$$\lim_{n \rightarrow \infty} P\{\xi_n < x\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt \quad (1.7.1)$$

(1) **切比雪夫大数定律** 设 X_1, X_2, \dots 是两两不相关 (相互独立) 的随机变量序列, 他们都有有限方差, 并且方差有共同的上界, 即 $D(X_i) \leq C, i = 1, 2, \dots$, 则对 $\forall \epsilon > 0$,

$$\lim_{n \rightarrow \infty} P \left\{ \left| \frac{1}{n} \sum_{i=1}^n X_i - \frac{1}{n} E(X_i) \right| < \epsilon \right\} = 1 \quad (1.7.2)$$

(2) **独立同分布下的大数定律** 设 X_1, X_2, \dots 是独立同分布的随机变量序列, 且 $E(X_i) = \mu, D(X_i) = \sigma^2, i = 1, 2, \dots$, 则对任意给 $\epsilon > 0$

$$\lim_{n \rightarrow \infty} P \left\{ \left| \frac{1}{n} \sum_{i=1}^n X_i - \mu \right| < \epsilon \right\} = 1 \quad (1.7.3)$$

(3) **伯努利大数定律** 设 μ_n 是 n 重伯努利实验中事件 A 发生的次数, p 是事件 A 发生的概率, 引入

$$X_i = \begin{cases} 1, & \text{如果第 } i \text{ 次实验 } A \text{ 发生} \\ 0, & \text{其他} \end{cases}$$

则

$$\mu_n = \sum_{i=1}^n X_i$$

有

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

是事件 A 发生的频率。设 p 是事件 A 发生的概率, 则对任给的 $\epsilon > 0$, 有

$$\lim_{n \rightarrow \infty} P \left\{ \left| \frac{\mu_n}{n} - p \right| < \epsilon \right\} = 1 \quad (1.7.4)$$

$$\lim_{n \rightarrow \infty} P \left\{ \left| \frac{\mu_n}{n} - p \right| \geq \epsilon \right\} = 0 \quad (1.7.5)$$

表明, 当重复实验次数 n 充分大时, 事件 A 发生的频率 $\frac{\mu_n}{n}$ 与事件 A 的概率 p 有较大偏差的概率很小。

1.7.2 中心极限定理

(1) **随机变量序列的标准化随机变量** n 个随机变量序列 X_1, X_2, \dots, X_n 的标准化随机变量

$$Z_n = \frac{\sum_{k=1}^n X_k - E \left(\sum_{k=1}^n X_k \right)}{\sqrt{D \left(\sum_{k=1}^n X_k \right)}} \quad (1.7.6)$$

(2) **林德贝格—莱维定理** 设 X_1, X_2, \dots 是独立同分布的随机变量序列, 且 $E(X_i) = \mu$, $D(X_i) = \sigma^2$, $i = 1, 2, \dots$ 则

$$\lim_{n \rightarrow \infty} P \left\{ \frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} \leq x \right\} = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt \quad (1.7.7)$$

表明当 n 充分大时, n 个具有期望和方差的独立同分布随机变量之和近似服从正态分布。

(3) 棣莫弗—拉普拉斯极限定理 设 μ_n 是 n 次伯努利实验中事件 A 出现的次数, $0 < p < 1$, 则对任意有限区间 $[a, b]$

• 当

$$a \leq x_k = \frac{k - np}{\sqrt{npq}} \leq b$$

及 $n \rightarrow \infty$ 时, 一致有

$$\frac{P\{\mu_n = k\}}{\frac{1}{\sqrt{npq}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x_k^2}} \rightarrow 1 \quad (1.7.8)$$

• 当 $n \rightarrow \infty$ 时, 一致有

$$P\left\{a \leq \frac{\mu_n - np}{\sqrt{npq}} < b\right\} \rightarrow \int_a^b \phi_x dx \quad (1.7.9)$$

其中

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

第二章 数理统计的基本概念

2.1 总体与样本

(1) **总体与个体** 研究对象的全体称为**总体**，总体中所包含的个体的个数称为总体的**容量**。总体中每个成员称为**个体**。

- 由于每个个体的出现是随机的，所以相应的数量指标的出现也带有随机性。从而可以把这种数量指标看作一个随机变量 X ，因此随机变量 X 的分布就是该数量指标在总体中的分布。
- 总体就可以用一个随机变量及其分布来描述。因此在理论上可以把总体与概率分布等同起来。统计中，总体这个概念的要旨是：总体就是一个随机变量（向量）或一个概率分布。

(2) **样本** 总体中抽出若干个体而成的集体，称为**样本**。样本中所含个体的个数，称为**样本容量**。

(3) **抽样** 统计中，采用的抽样方法是随机抽样法，即子样中每个个体是从总体中随意地取出来的。

(i) 抽样的分类

- **重复(返回)抽样**: 从总体中抽取个体检查后放回, 总体成分不变(分布不变)。样本 X_1, X_2, \dots, X_n 相互独立, 与总体有相同的分布。
- **非重复(无返回)抽样**: 对有限总体取出样本后改变了总体的成分, 所以 X_1, X_2, \dots, X_n 不相互独立; 对无限总体而言做无返回抽取, 并不改变总体的成分, X_1, X_2, \dots, X_n 相互独立, 与总体有相同的分布。

(ii) 常用方法 简单随机抽样。

- 代表性 (随机性):。从总体中抽取样本的每一个分量 X_k 是随机的, 每一个个体被抽到的可能性相同。
- 独立同分布性: X_1, X_2, \dots, X_n 相互独立, 其中每一个分量 X_k 与所考察的总体有相同的分布。

(iii) **样本联合分布** 若总体的分布函数为 $F(x)$ 、概率密度为 $f(x)$, 则其简单随机样本的联合分布函数为

$$F_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n) = F(1)F(2) \cdots F(n) \quad (2.1.1)$$

其简单随机样本的联合概率密度函数为

$$f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n) = f(1)f(2) \cdots f(n) \quad (2.1.2)$$

(4) **样本经验分布函数** 在 n 次独立重复实验中, 事件 $\{X \leq x\}$ 发生的频率

$$\hat{F}_n(x; X_1, X_2, \dots, X_n) = \frac{1}{n} \sum_{i=1}^n I(X_i \leq x) \quad (2.1.3)$$

具有分布函数的一切性质。是在每个数据点 X_i 上权重相等的均匀分布的分布函数。

(i) 性质

- 给定 x , $\hat{F}_n(x)$ 是一个随机变量: $n\hat{F}_n(x)$ 服从二项分布 $b(n, F(x))$
- $E(\hat{F}_n(x)) = F(x)$
- $D(\hat{F}_n(x)) = \frac{F(x)(1-F(x))}{n} \rightarrow 0$
- $\hat{F}_n(x) \xrightarrow{P} F(x)$
- Dvoretzky-Kiefer-Wolfowitz (DKW) 不等式: 如果 $X_1, X_2, \dots, X_n \sim F$, 则对任意 $\epsilon > 0$

$$P\left\{\sup_x |\hat{F}_n(x) - F(x)| > \epsilon\right\} \leq 2e^{-2n\epsilon^2} \quad (2.1.4)$$

(ii) 格列汶科定理 当 $n \rightarrow \infty$ 时, $\hat{F}_n(x)$ 以概率 1 关于 x 一致收敛于 $F(x)$, 即

$$P\left\{\lim_{n \rightarrow \infty} \sup_{-\infty < x < \infty} |\hat{F}_n(x) - F(x)| = 0\right\} = 1 \quad (2.1.5)$$

当样本容量 n 足够大时, 对所有的 x , $\hat{F}_n(x)$ 与 $F(x)$ 之差的绝对值都很小, 这件事发生的概率为 1。

2.2 统计量

(1) 统计量 不含任何未知参数的样本的函数称为统计量。它是完全由样本决定的量。

- 统计量是随机变量
- 设 X_1, X_2, \dots, X_n 是来自总体 X 的一个样本, x_1, x_2, \dots, x_n 是一个样本的观察值, 则 $g(x_1, x_2, \dots, x_n)$ 也是统计量 $g(X_1, X_2, \dots, X_n)$ 的观察值。

(2) 常见统计量

- 样本均值:

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

- 样本方差:

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

- 样本标准差:

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

- 样本 k 阶原点矩:

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

- 样本 k 阶中心矩:

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

- 顺序统计量: 设 (x_1, x_2, \dots, x_n) 为样本 (X_1, X_2, \dots, X_n) 的一个观察值, 且 $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ 。当 (X_1, X_2, \dots, X_n) 取值为 (x_1, x_2, \dots, x_n) 时, 定义随机变量 $X_{(k)} = x_{(k)}$ 则称统计量 $(X_{(1)}, X_{(2)}, \dots, X_{(n)})$ 为顺序统计量。

- 极差:

$$D_n = X_{(n)} - X_{(1)}$$

(3) 常见统计量的性质

- $E(\bar{X}) = E(X) = \mu$
- $D(\bar{X}) = \frac{D(X)}{n} = \frac{\sigma^2}{n}$
- $E(S^2) = D(X) = \sigma^2$
- 若总体 k 阶矩 $E(X^k) = \mu_k$ 存在, 则当 $n \rightarrow \infty$ 时,

$$A_k = \frac{1}{n} \sum_{i=1}^n X_i^k \xrightarrow{P} \mu_k \quad (2.2.1)$$

推广

$$g(A_1, A_2, \dots, A_n) \xrightarrow{P} g(\mu_1, \mu_2, \dots, \mu_n) \quad (2.2.2)$$

(4) 统计量的分布 设总体分布的特征函数为 $\phi_X(t)$, 则 \bar{X} 的特征函数是

$$\phi_{\bar{X}}(t) = \left(\phi_X\left(\frac{t}{n}\right) \right)^n \quad (2.2.3)$$

- 正态分布: $\bar{X} \sim N(\mu, \sigma^2/n)$

$$\phi_{\bar{X}}(t) = e^{i\mu t - \frac{1}{2} \frac{\sigma^2}{n} t^2}$$

- 泊松分布:

$$\phi_{\bar{X}}(t) = \exp(n\lambda(e^{it/n} - 1))$$

- 指数分布: $\bar{X} \sim \Gamma(n-1, \frac{1}{n\lambda})$

$$\phi_{\bar{X}}(t) = \left(1 - \frac{it}{n\lambda}\right)$$

- 二项分布:

$$\phi_{\bar{X}}(t) = (pe^{it/n} + q)^{Nn}$$

2.3 三大统计抽样分布

(1) χ^2 分布 设 X_1, X_2, \dots, X_n 互相独立, 都服从正态分布 $N(0, 1)$ 则称随机变量

$$\chi^2 = X_1^2 + X_2^2 + \dots + X_n^2 \quad (2.3.1)$$

所服从的分布为自由度为 n 的 χ^2 分布。(i) 密度函数 χ^2 分布的密度函数为

$$f(x; n) = \begin{cases} \frac{1}{2^{\frac{n}{2}} \Gamma(n/2)} x^{\frac{n}{2}-1} e^{-\frac{x}{2}} & x \geq 0 \\ 0 & x < 0 \end{cases} \quad (2.3.2)$$

(ii) 期望和方差 $E(X) = n, D(X) = 2n$ 。

(iii) 性质

- 设 X_1, X_2, \dots, X_n 互相独立, 都服从正态分布 $N(0, 1)$ 则

$$\chi^2 = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \mu)^2 \sim \chi^2(n) \quad (2.3.3)$$

- 可加性: 设 $X_1 \sim \chi^2(n_1), X_2 \sim \chi^2(n_2)$ 且 X_1, X_2 互相独立, 则

$$X_1 + X_2 \sim \chi^2(n_1 + n_2) \quad (2.3.4)$$

- 若 $\chi^2 \sim \chi^2(n)$ 则当 n 充分大时, $\frac{X-n}{\sqrt{2n}}$ 的分布近似正态分布 $N(0, 1)$ 。

(iv) **上分位点** 对于给定的正数 $\alpha (0 < \alpha < 1)$ 称满足条件

$$P\{\chi^2 > \chi_\alpha^2(n)\} = \int_{\chi_\alpha^2(n)}^{\infty} f(y)dy = \alpha \quad (2.3.5)$$

的点 $\chi_\alpha^2(n)$ 称为 $\chi^2(n)$ 分布的上 α 分位点。

(2) **t 分布** 设 $X \sim N(0,1), Y \sim \chi^2(n)$ 且 X 与 Y 相互独立, 则称变量

$$T = \frac{X}{\sqrt{\frac{Y}{n}}} \quad (2.3.6)$$

所服从的分布为自由度为 n 的 t 分布, 记为 $T \sim t(n)$ 。

(i) **密度函数** t 分布的密度函数为

$$h(t) = \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}} \quad -\infty < t < \infty \quad (2.3.7)$$

(ii) **期望和方差** $E(t) = 0, D(t) = \frac{n}{n-2}$ 。

(iii) **性质**

- 密度函数关于 $t = 0$ 对称, 当 n 充分大时, 其图形近似于标准正态分布概率密度的图形。且

$$\lim_{n \rightarrow \infty} h(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \quad (2.3.8)$$

即当 n 足够大时, $T \sim N(0,1)$ 。

(3) **F 分布** 设 $X \sim \chi^2(n_1), Y \sim \chi^2(n_2)$ 且 X, Y 独立, 则称统计量

$$F = \frac{X/n_1}{Y/n_2} \quad (2.3.9)$$

服从自由度为 n_1, n_2 的 F 分布, n_1 为第一自由度, n_2 为第二自由度, 记为 $F \sim F(n_1, n_2)$ 。

(i) **密度函数** F 分布的密度函数为

$$f(y) = \begin{cases} \frac{\Gamma(\frac{n_1+n_2}{2})}{\Gamma(\frac{n_1}{2})\Gamma(\frac{n_2}{2})} \left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}} (y)^{\frac{n_1}{2}-1} \left(1 + \frac{n_1}{n_2}y\right)^{-\frac{n_1+n_2}{2}} & y > 0 \\ 0 & y \leq 0 \end{cases} \quad (2.3.10)$$

(ii) **期望和方差** $E(t) = \frac{n_2}{n_2-2}$ 。

(iii) **性质**

- 即它的数学期望并不依赖于第一自由度。
- 若 $F \sim F(n_1, n_2)$ 则 $\frac{1}{F} \sim F(n_2, n_1)$
- $F_{1-\alpha}(n_1, n_2) = 1/F_\alpha(n_2, n_1)$

2.4 正态总体抽样分布定理

设总体 X 的均值为 μ , 方差为 σ^2 , X_1, X_2, \dots, X_n 是来自总体的一个样本, 则样本均值 \bar{X} 和样本方差 S^2 有

$$E(\bar{X}) = \mu \quad (2.4.1)$$

$$D(\bar{X}) = \frac{\sigma^2}{n} \quad (2.4.2)$$

$$E(S^2) = \sigma^2 \quad (2.4.3)$$

(1) **样本均值的分布** 设 X_1, X_2, \dots, X_n 是来自正态总体 $N(\mu, \sigma^2)$ 的样本, \bar{X} 是样本均值, 则有

$$X \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad (2.4.4)$$

则

$$\frac{\bar{X} - \mu}{\frac{\sigma^2}{\sqrt{n}}} \sim N(0, 1) \quad (2.4.5)$$

(2) **样本方差的分布** 设 X_1, X_2, \dots, X_n 是来自正态总体 $N(\mu, \sigma^2)$ 的样本, \bar{X} 是样本均值, S^2 是样本方差, 则有

•

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1) \quad (2.4.6)$$

• \bar{X} 与 S^2 独立

(3) **样本均值的分布** 设 X_1, X_2, \dots, X_n 是来自正态总体 $N(\mu, \sigma^2)$ 的样本, \bar{X} 是样本均值, S^2 是样本方差, 则有

$$\frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t(n-1) \quad (2.4.7)$$

(4) **两总体样本均值差的分布** 设 $X \sim N(\mu_1, \sigma_1^2), Y \sim N(\mu_2, \sigma_2^2)$ 且 X, Y 独立, X_1, X_2, \dots, X_n 是来自正态总体 X 的样本, Y_1, Y_2, \dots, Y_n 是来自正态总体 Y 的样本, \bar{X}, \bar{Y} 分别是这两个样本的均值, S_1^2, S_2^2 分别是这两个样本的方差, 则有

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1-1, n_2-1) \quad (2.4.8)$$

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1 + n_2 - 2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2) \quad (2.4.9)$$

(5) **Cochran 定理** 设 X_1, X_2, \dots, X_n 独立同分布, 服从 $N(0, 1)$, $X = (X_1, X_2, \dots, X_n)^T$,

$$Q = \sum_{i=1}^n X_i^2 = \sum_{l=1}^n Q_l$$

是 $X = (X_1, X_2, \dots, X_n)^T$ 的二次型, $Q = X^T A_l X$ 则下述结论成立的充要条件是

$$n = \sum_{i=1}^k n_i$$

- Q_1, Q_2, \dots, Q_k 相互独立
- $Q_l \sim \chi^2(n_l)$

第三章 参数估计

参数估计 设有一个统计总体，总体的分布函数为 $F(x; \theta)$ 其中 θ 是位置参数（可以是向量），现在从总体抽样，得到样本 X_1, X_2, \dots, X_n 要根据改样本对参数 θ 做出估计，或估计 θ 的某个函数 $g(\theta)$ 。这类问题被称为参数估计。

3.1 点估计

3.1.1 点估计的方法

(1) **点估计** 我们需要构造出适当的样本的函数 $T(X_1, X_2, \dots, X_n)$ ，每当有了样本，就代入该函数中算出一个值，用来作为 θ 的估计值。 $T(X_1, X_2, \dots, X_n)$ 称为参数 θ 的点估计量。把样本值代入 $T(X_1, X_2, \dots, X_n)$ 中，得到 θ 的一个点估计值。

(2) **矩估计** 用样本原点矩估计相应的总体原点矩，又用样本原点矩的连续函数估计相应的总体原点矩的连续函数，这种参数点估计法称为矩估计法。

(i) **做法** 设总体的分布函数中含有 k 个未知参数 $\theta_1, \theta_2, \dots, \theta_n$ 那么他的前 k 阶矩 $\mu_1, \mu_2, \dots, \mu_n$ 一般都是这 k 个参数的函数，记为

$$\mu_i = \mu_i(\theta_1, \theta_2, \dots, \theta_n), \quad i = 1, 2, \dots, k$$

从这 k 个方程中解出

$$\theta_i = \theta_j(\mu_1, \mu_2, \dots, \mu_n)$$

用各 μ_i 的估计量 A_i 分别代替上式中的各个 μ_i 即可得到 θ_i 的矩估计量

$$\hat{\theta}_i = \theta_j(A_1, A_2, \dots, A_n)$$

矩估计量的观察值称为矩估计值。

(ii) **优点** 简单易行，并不需要事先知道总体是什么分布。

(iii) **缺点** 当总体类型已知时，没有充分利用分布提供的信息。一般场合下，矩估计量不具有唯一性。矩估计不唯一时处理的原则是尽量使用低阶矩。

(3) **最大似然法** 再一次抽样中，若得到观测值 x_1, x_2, \dots, x_n 则选取 $\hat{\theta}(x_1, x_2, \dots, x_n)$ 作为 θ 的估计值，使得当 $\theta = \hat{\theta}$ 时，样本出现的概率最大。

(i) **做法** 设 x_1, x_2, \dots, x_n 是 X_1, X_2, \dots, X_n 的一个样本值， X 的联合分布率是

$$\prod_{i=1}^n p(x_i; \theta), \theta \in \Theta$$

事件 $P\{X_1 = x_1, X_2 = x_2, \dots, X_n = x_n\}$ 发生的概率为

$$\begin{aligned} L(\theta) &= L(x_1, x_2, \dots, x_n; \theta) = \prod_{i=1}^n p(x_i; \theta) \\ L(\theta) &= L(x_1, x_2, \dots, x_n; \theta) = \prod_{i=1}^n f(x_i; \theta) \end{aligned} \quad (3.1.1)$$

式 (3.1.1) 称为样本似然函数。取 $\hat{\theta}$ 使得

$$L(x_1, x_2, \dots, x_n; \hat{\theta}) = \max_{\theta \in \Theta} L(x_1, x_2, \dots, x_n; \theta)$$

一般可由下式求得:

$$\frac{dL(\theta)}{d\theta} = 0 \quad (3.1.2)$$

或

$$\frac{d \ln L(\theta)}{d\theta} = 0 \quad (3.1.3)$$

若又多个参数, 可令

$$\frac{dL(\theta)}{d\theta_i} = 0 \quad \frac{d \ln L(\theta)}{d\theta_i} = 0 \quad (3.1.4)$$

求解 k 个方程组求得 $\theta_1, \theta_2, \dots, \theta_n$ 的最大似然估计值。

3.1.2 估计量的评选标准

(1) **无偏性** 设 $\hat{\theta}(X_1, X_2, \dots, X_n)$ 是未知参数 θ 的估计量, 若

$$E(\hat{\theta}) = \theta \quad (3.1.5)$$

则称 $\hat{\theta}$ 为 θ 的无偏估计。

(2) **有效性** 设 $\hat{\theta}_1 = \hat{\theta}_1(X_1, X_2, \dots, X_n), \hat{\theta}_2 = \hat{\theta}_2(X_1, X_2, \dots, X_n)$ 均是 θ 的无偏估计, 若

$$D(\hat{\theta}_1) \leq D(\hat{\theta}_2) \quad (3.1.6)$$

则称 $\hat{\theta}_1$ 比 $\hat{\theta}_2$ 有效。

(3) **最小方差无偏估计** 如果存在 θ 的无偏估计量 $\hat{\theta}_0$ 使得对于 θ 的任一方差存在的无偏估计量 $\hat{\theta}$ 都有

$$D(\hat{\theta}_0) \leq D(\hat{\theta}) \quad (3.1.7)$$

则称 $\hat{\theta}_0$ 是 $\hat{\theta}$ 的最小方差无偏估计量, 缩写为 UMVUE。最小方差无偏估计是一种最优估计。基于充分完备统计量的无偏估计一定是 UMVUE。

(4) **优效估计** 若无偏估计 $\hat{\theta}_0$ 的方差 $D(\hat{\theta}_0) = I_R$ 则称 $\hat{\theta}_0$ 为 θ 的优效估计。

(i) **Cramer-Rao 不等式** 设总体 $X \sim f(x; \theta)$, θ 是未知参数, X_1, X_2, \dots, X_n 是 X 的一个简单样本, $\hat{\theta} = \hat{\theta}(X_1, X_2, \dots, X_n)$ 是 θ 的无偏估计, 则

$$D(\hat{\theta}) \geq \frac{1}{nI(\theta)} \quad (3.1.8)$$

右端称为 C-R 下界, 记为 I_R 。有时能找到无偏估计使它的方差达到这个下界, 有时达不到。

(ii) **Fisher 信息量** 设总体 X 的概率密度函数为 $f(x; \theta), \theta \in \Theta$ 且满足下列条件 (称为正则条件), 则称

$$I(\theta) = E \left(\frac{\partial \ln f(X; \theta)}{\partial \theta} \right)^2 \quad (3.1.9)$$

为总体分布的 Fisher 信息量。

- Θ 是实数轴上的一个开区间;
- 支撑 $S = \{x | f(x; \theta) > 0\}$ 与 θ 无关;
- $\frac{\partial f(x; \theta)}{\partial \theta}$ 存在且对 Θ 中一切 θ 有

$$\frac{\partial}{\partial \theta} \int_{-\infty}^{\infty} f(x; \theta) dx = \int_{-\infty}^{\infty} \frac{\partial}{\partial \theta} f(x; \theta) dx \quad (3.1.10)$$

- $E \left(\frac{\partial \ln f(X; \theta)}{\partial \theta} \right)^2$ 存在。

(iii) **Fisher 信息矩阵** 若 θ 为向量, 则称

$$I(\theta) = E \left[\left(\frac{\partial f(X; \theta)}{\partial \theta} \right) \left(\frac{\partial f(X; \theta)}{\partial \theta} \right)^T \right] \quad (3.1.11)$$

为总体分布的 Fisher 信息矩阵。

(iv) **Fisher 信息量的另一种表达式** 若 θ 为标量, $\frac{\partial^2 \ln f(X; \theta)}{\partial \theta^2}$ 存在且满足正则条件, 则

$$I(\theta) = -E \left(\frac{\partial^2 \ln f(X; \theta)}{\partial \theta^2} \right) \quad (3.1.12)$$

若 θ 为向量, $\frac{\partial^2 \ln f(X; \theta)}{\partial \theta \partial \theta^T}$ 存在且满足正则条件, 则

$$I(\theta) = -E \left(\frac{\partial^2 \ln f(X; \theta)}{\partial \theta \partial \theta^T} \right) \quad (3.1.13)$$

(5) **相合性 (一致性)** 设 $\hat{\theta}(X_1, X_2, \dots, X_n)$ 是参数 θ 的估计量, 若对于任意 $\theta \in \Theta$ 当 $n \rightarrow \infty$ 时 $\hat{\theta}(X_1, X_2, \dots, X_n)$ 依概率收敛到 θ , 则称 $\hat{\theta}$ 为 θ 的相合估计量。即对于任意 $\epsilon > 0$ 有

$$\lim_{n \rightarrow \infty} P\{|\hat{\theta} - \theta| < \epsilon\} = 1, \theta \in \Theta \quad (3.1.14)$$

3.2 区间估计

3.2.1 置信区间

(1) 置信区间

(i) **双侧置信区间** 设 θ 是一个待估参数, 给定 $\alpha > 0$ 若由样本 X_1, X_2, \dots, X_n 确定的两个统计量,

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

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

满足

$$P\{\underline{\theta} < \theta < \bar{\theta}\} = 1 - \alpha \quad (3.2.1)$$

则称区间 $(\underline{\theta}, \bar{\theta})$ 是 θ 置信水平 (置信度) 为 $1 - \alpha$ 的置信区间。 $\underline{\theta}$ 和 $\bar{\theta}$ 分别称为置信下限和置信上限。

(ii) 右侧置信区间 设 θ 是一个待估参数, 给定 $\alpha > 0$ 若由样本 X_1, X_2, \dots, X_n 确定的两个统计量,

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

满足

$$P\{\underline{\theta} < \theta\} = 1 - \alpha \quad (3.2.2)$$

则称区间 $(\underline{\theta}, +\infty)$ 是 θ 置信水平 (置信度) 为 $1 - \alpha$ 的右侧置信区间。 $\underline{\theta}$ 称为单侧置信下限。

(iii) 左侧置信区间 设 θ 是一个待估参数, 给定 $\alpha > 0$ 若由样本 X_1, X_2, \dots, X_n 确定的两个统计量,

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

满足

$$P\{\theta < \bar{\theta}\} = 1 - \alpha \quad (3.2.3)$$

则称区间 $(-\infty, \bar{\theta})$ 是 θ 置信水平 (置信度) 为 $1 - \alpha$ 的左侧置信区间。 $\bar{\theta}$ 称为单侧置信上限。

(2) 求法

- 寻找参数 θ 的一个良好的点估计量 $T(X_1, X_2, \dots, X_n)$ 。
- 寻找参数 θ 和估计量 T 的函数 $U(T, \theta)$ 且其分布已知。
- 对于给定置信水平 $1 - \alpha$ 根据 $U(T, \theta)$ 的分布, 确定常数 a, b 使得

$$P\{a < U(T, \theta) < b\} = 1 - \alpha$$

- 对 $a < U(T, \theta) < b$ 作等价变形, 得到形式 $\underline{\theta} < \theta < \bar{\theta}$ 。

3.2.2 正态总体均值和方差的区间估计

表 3.1 正态总体均值和方差的区间估计

待估参数	条件	统计量	置信区间
μ	已知 σ^2	$U = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$	$\left[\bar{X} - \frac{\sigma}{\sqrt{n}} u_{\frac{\alpha}{2}}, \bar{X} + \frac{\sigma}{\sqrt{n}} u_{\frac{\alpha}{2}} \right]$
μ	未知 σ^2	$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$	$\left[\bar{X} - \frac{S}{\sqrt{n}} t_{\frac{\alpha}{2}}(n-1), \bar{X} + \frac{S}{\sqrt{n}} t_{\frac{\alpha}{2}}(n-1) \right]$
σ^2	已知 μ	$\chi^2 = \frac{\sum_{i=1}^n (X_i - \mu)^2}{\sigma^2} \sim \chi^2(n)$	$\left[\frac{\sum_{i=1}^n (X_i - \mu)^2}{\chi_{\frac{\alpha}{2}}^2(n)}, \frac{\sum_{i=1}^n (X_i - \mu)^2}{\chi_{1-\frac{\alpha}{2}}^2(n)} \right]$
σ^2	未知 μ	$\chi^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$	$\left[\frac{(n-1)S^2}{\chi_{\frac{\alpha}{2}}^2(n)}, \frac{(n-1)S^2}{\chi_{1-\frac{\alpha}{2}}^2(n)} \right]$

(续表)

待估参数	条件	统计量	置信区间
$\mu_1 - \mu_2$	已知 σ_1, σ_2	$U = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$	$\left[(\bar{X} - \bar{Y}) - u_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \right. \\ \left. (\bar{X} - \bar{Y}) + u_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right]$
$\mu_1 - \mu_2$	$\sigma_1 = \sigma_2 = \sigma$ σ 未知	$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_w \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$	$\left[(\bar{X} - \bar{Y}) - t_{\frac{\alpha}{2}} S_w \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \right. \\ \left. (\bar{X} - \bar{Y}) + t_{\frac{\alpha}{2}} S_w \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right]$
$\frac{\sigma_1}{\sigma_2}$	已知 μ_1, μ_2	$F = \frac{\frac{\frac{1}{n_1} \sum_{i=1}^{n_1} (X_i - \mu_1)^2}{\frac{1}{n_2} \sum_{i=1}^{n_2} (Y_i - \mu_2)^2}}{\frac{\sigma^2}{\sigma^2}} \sim F(n_1, n_2)$	$\left[\frac{\frac{1}{n_1} \sum_{i=1}^{n_1} (X_i - \mu_1)^2}{\frac{1}{n_2} \sum_{i=1}^{n_2} (Y_i - \mu_2)^2} \frac{1}{F_{\frac{\alpha}{2}}(n_1, n_2)}, \right. \\ \left. \frac{\frac{1}{n_1} \sum_{i=1}^{n_1} (X_i - \mu_1)^2}{\frac{1}{n_2} \sum_{i=1}^{n_2} (Y_i - \mu_2)^2} \frac{1}{F_{1-\frac{\alpha}{2}}(n_1, n_2)} \right]$
$\frac{\sigma_1}{\sigma_2}$	已知 μ_1, μ_2	$F = \frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}} \sim F(n_1 - 1, n_2 - 1)$	$\left[\frac{S_1^2}{S_2^2} \frac{1}{F_{\frac{\alpha}{2}}(n_1 - 1, n_2 - 2)}, \right. \\ \left. \frac{S_1^2}{S_2^2} \frac{1}{F_{1-\frac{\alpha}{2}}(n_1 - 1, n_2 - 2)} \right]$

第四章 假设检验

4.1 假设检验

- (1) **参数假设检验** 总体的分布类型已知，对未知参数作出假设，用总体中的样本检验此项假设是否成立，就称为参数假设检验。
- (2) **非参数假设检验** 对总体分布函数的形式作出假设，用总体中的样本检验此项假设是否成立，就称为非参数假设检验。
- (3) **基本步骤**
- 题出原假设 H_0 ，确定备择假设 H_1 ；
 - 构造分布已知的合适的统计量；
 - 由给定的检验水平 α ，求出在 H_0 成立的条件下的临界值；
 - 计算统计量的样本观测值，如果落在拒绝域内，则拒绝原假设；否则，接收原假设。
- (4) **假设检验的两类错误**
- (i) **第一类错误（弃真错误）** 原假设为真，而检验结果拒绝原假设。记为 α 。
- (ii) **第二类错误（取伪错误）** 原假设为假，而检验结果接受原假设。记为 β 。
- (iii) **原则**
- 在限制 α 的前提下，使 β 尽可能小。
 - 往往把不轻易否定的命题作为原假设
 - 通常控制犯第一类错误的概率
- (5) **显著性检验** 只对犯第一类错误的概率加以控制，而不考虑犯第二类错误的概率。即求

$$P\{\text{拒绝}H_0 | H_0\text{为真}\} \leq \alpha$$

称 α 为显著性水平。

4.2 正态总体参数的假设检验

表 4.1 正态总体均值和方差的假设检验

检验问题	原假设	备择假设	检验统计量	拒绝域
μ 检验 σ^2 已知	$\mu = \mu_0$	$\mu \neq \mu_0$	$U = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$	$ U \geq z_{\frac{\alpha}{2}}$
	$\mu \geq \mu_0$	$\mu < \mu_0$		$U \leq -z_\alpha$
	$\mu \leq \mu_0$	$\mu > \mu_0$		$U \geq z_\alpha$

(续表)

检验问题	原假设	备择假设	检验统计量	拒绝域
μ 检验 σ^2 未知	$\mu = \mu_0$	$\mu \neq \mu_0$	$T = \frac{\bar{X} - \mu_0}{\frac{S}{\sqrt{n}}} \sim t(n-1)$	$ T \geq t_{\frac{\alpha}{2}}(n-1)$
	$\mu \geq \mu_0$	$\mu < \mu_0$		$T \leq t_{\alpha}(n-1)$
	$\mu \leq \mu_0$	$\mu > \mu_0$		$T \geq t_{\alpha}(n-1)$
σ^2 检验 μ 已知	$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$	$\chi^2 = \frac{\sum_{i=1}^n (X_i - \mu)^2}{\sigma_0^2} \sim \chi^2(n)$	$\chi^2 \leq \chi_{1-\frac{\alpha}{2}}^2(n)$, 或 $\chi^2 \geq \chi_{\frac{\alpha}{2}}^2(n)$
	$\sigma^2 \geq \sigma_0^2$	$\sigma^2 < \sigma_0^2$		$\chi^2 \leq \chi_{1-\alpha}^2(n)$
	$\sigma^2 \leq \sigma_0^2$	$\sigma^2 > \sigma_0^2$		$\chi^2 \geq \chi_{\alpha}^2(n)$
σ^2 检验 μ 已知	$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$	$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2(n-1)$	$\chi^2 \leq \chi_{1-\frac{\alpha}{2}}^2(n-1)$ 或 $\chi^2 \geq \chi_{\frac{\alpha}{2}}^2(n-1)$
	$\sigma^2 \geq \sigma_0^2$	$\sigma^2 < \sigma_0^2$		$\chi^2 \leq \chi_{1-\alpha}^2(n-1)$
	$\sigma^2 \leq \sigma_0^2$	$\sigma^2 > \sigma_0^2$		$\chi^2 \geq \chi_{\alpha}^2(n-1)$
$\mu_1 - \mu_2$ 检验 σ_1, σ_2 已知	$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	$U = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$	$ U \geq z_{\frac{\alpha}{2}}$
	$\mu_1 = \mu_2$	$\mu_1 < \mu_2$		$U \leq -z_{\alpha}$
	$\mu_1 = \mu_2$	$\mu_1 > \mu_2$		$U \geq z_{\alpha}$
$\mu_1 - \mu_2$ 检验 σ_1, σ_2 未知	$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} S_w} \sim T(n_1 + n_2 - 2)$	$ T \geq z_{\frac{\alpha}{2}}$
	$\mu_1 = \mu_2$	$\mu_1 < \mu_2$		$T \leq -z_{\alpha}$
	$\mu_1 = \mu_2$	$\mu_1 > \mu_2$		$T \geq z_{\alpha}$
$\frac{\sigma_1}{\sigma_2}$ 检验 μ_1, μ_2 已知	$\sigma_1 = \sigma_2$	$\sigma_1 \neq \sigma_2$	$F = \frac{\frac{1}{n_1} \sum_{i=1}^{n_1} (X_i - \mu_1)^2}{\frac{1}{n_2} \sum_{i=1}^{n_2} (Y_i - \mu_2)^2} \sim F(n_1, n_2)$	$F \leq F_{1-\frac{\alpha}{2}}(n_1, n_2)$ 或 $F \geq F_{\frac{\alpha}{2}}(n_1, n_2)$
	$\sigma_1 = \sigma_2$	$\sigma_1 < \sigma_2$		$F \leq F_{1-\alpha}(n_1, n_2)$
	$\sigma_1 = \sigma_2$	$\sigma_1 > \sigma_2$		$F \geq F_{\alpha}(n_1, n_2)$
$\frac{\sigma_1}{\sigma_2}$ 检验 μ_1, μ_2 未知	$\sigma_1 = \sigma_2$	$\sigma_1 \neq \sigma_2$	$F = \frac{S_1^2}{S_2^2} \sim F(n_1 - 1, n_2 - 1)$	$F \leq F_{1-\frac{\alpha}{2}}(n_1 - 1, n_2 - 1)$ 或 $F \geq F_{\frac{\alpha}{2}}(n_1 - 1, n_2 - 1)$
	$\sigma_1 = \sigma_2$	$\sigma_1 < \sigma_2$		$F \leq F_{1-\alpha}(n_1 - 1, n_2 - 1)$
	$\sigma_1 = \sigma_2$	$\sigma_1 > \sigma_2$		$F \geq F_{\alpha}(n_1 - 1, n_2 - 1)$

4.3 似然比检验

(1) **似然比检验** 假设总体的密度函数或概率分布为 $f(x; \theta)$ ，其中 $\theta = (\theta_1, \theta_2, \dots, \theta_n)$ 是 k 维参数， $\theta \in \Theta$ 。检验其中的 r 个参数，有原假设

$$H_0 = \theta_1 = \theta_{10}, \theta_2 = \theta_{20}, \dots, \theta_r = \theta_{r0}$$

其中 $\theta_{10}, \theta_{20}, \theta_{r0}$ 是制定的常数。似然比定义为

$$LR = \frac{\sup_{H_0} L(\theta)}{\sup_{\theta \in \Theta} L(\theta)} = \frac{L(\tilde{\theta})}{L(\hat{\theta})} \quad (4.3.1)$$

拒绝域为 $C_1 = \{LR < c\}$, $P\{C_1|H_0\} \leq \alpha$ 。

- $0 < LR \leq 1$
- $\hat{\theta} \xrightarrow{P} \theta$

(2) **定理** 当似然函数 $L(\theta)$ 为三阶可导函数，且 $\int_{-\infty}^{\infty} \ln f(x; \theta) dx$ 连续可导，则有： $-2 \ln LR$ 的极限分布为 $\chi^2(r), (n \rightarrow \infty)$ ，拒绝域为 $C_1 = \{\chi^2 = -2 \ln LR > \chi_{\alpha}^2(r)\}$ 。

(3) **Newman-Pearson 引理** 原假设和备择假设都是简单假设时，似然比检验是一致最佳检验，即由似然比检验法导出的拒绝域，是在同一显著性水平下犯第二类错误最小的拒绝域。

(4) **样本容量 n 的确定** 原假设和备择假设都是简单假设（即参数只取参数空间的一个点）时，寻找最小的样本容量，使得两类错误的概率控制在预制范围内。

- 总体方差已知，正态总体均值的右边检验：

$$n = (\mu_{\alpha} + \mu_{\beta})^2 \frac{\sigma^2}{(\mu_1 - \mu_0)^2} \quad (4.3.2)$$

- 总体方差未知，正态总体均值的右边检验：

$$n = (t_{\alpha}(n-1) + t_{\beta}(n-1))^2 \frac{S^2}{(\mu_1 - \mu_0)^2} \quad (4.3.3)$$

- 总体期望未知，正态总体方差的右边检验：

$$\chi_{1-\beta}^2(n-1) = \frac{\sigma_0^2}{\sigma_1^2}(n-1) \quad (4.3.4)$$

4.4 非参数假设检验

4.4.1 分布拟合检验

(1) **分布拟合检验** 设总体 X 的实际分布函数为 $F(x)$ ，它是未知的。 X_1, X_2, \dots, X_n 为来自总体 X 的样本。根据这个样本来检验总体 X 的分布函数 $F(x)$ 是否等于某个给定的分布函数 $F_0(x)$ ，即检验假设

$$H_0 : F(x) = F_0(x)$$

$$H_1 : F(x) \neq F_0(x)$$

- 若总体 X 为离散型的，则 H_0 相当于总体 X 的分布律为 $P\{X = x_i\}, i = 1, 2, \dots$ 。
- 若总体 X 为连续型的，则 H_0 相当于总体 X 的概率密度为 $f(x)$ 。

- (2) **$F(x)$ 不含未知参数** 记 Ω 为 X 的所有可能取值的全体, 将 Ω 分为 k 个两两互不相交的子集 A_1, A_2, \dots, A_k 。以 $f_i (i = 1, 2, \dots, k)$ 表示样本观察值 x_1, x_2, \dots, x_n 中落入 A_1, A_2, \dots, A_k 的个数。当 H_0 为真且 n 充分大时, 统计量

$$\chi^2 = \sum_{i=1}^k \frac{(f_i - np_i)^2}{np_i} = \sum_{i=1}^k \frac{f_i^2}{np_i} - n \quad (4.4.1)$$

近似服从 $\chi^2(k-1)$ 分布。给定显著性水平, 则拒绝域为 $\chi^2 \geq \chi_{\alpha}^2(k-1)$ 。

- (3) **$F(x)$ 含有未知参数** 此时, 首先在假设下利用样本求出未知参数的最大似然估计, 以估计值作为参数值, 然后再根据 H_0 中所假设的 X 的分布函数 $F(x)$ 求出 p_i 的估计值。统计量

$$\chi^2 = \sum_{i=1}^k \frac{(f_i - np_i)^2}{np_i} = \sum_{i=1}^k \frac{f_i^2}{np_i} - n$$

近似服从 $\chi^2(k-r-1)$ 分布, 其中 r 是 X 的分布函数 $F(x)$ 包含的未知参数的个数。给定显著性水平, 则拒绝域为 $\chi^2 \geq \chi_{\alpha}^2(k-r-1)$ 。

- (4) **样本数据的分类** 运用 χ^2 检验法检验总体分布, 把样本数据进行分类时

- 大样本, 通常取 $n \geq 50$ 。
- 各组理论频数 $np_i \geq 5$ 或 $n\hat{p}_i \geq 5$ 。
- 一般把数据分成 7 到 14 组, 有时为了保证各组 $np_i \geq 5$ 组数可以少于 7 组。

4.4.2 独立性检验

- (1) **列联表** $r \times s$ 列联表

	B_1	B_2	\dots	B_s	总和
A_1	n_{11}	n_{12}	\dots	n_{1s}	$n_{1\cdot}$
A_2	n_{21}	n_{22}	\dots	n_{2s}	$n_{2\cdot}$
\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
A_r	n_{r1}	n_{r2}	\dots	n_{rs}	$n_{r\cdot}$
总和	$n_{\cdot 1}$	$n_{\cdot 2}$	\dots	$n_{\cdot s}$	$n_{\cdot \cdot}$

其中

$$n_{i\cdot} = \sum_{j=1}^s n_{ij}, i = 1, 2, \dots, r \quad n_{\cdot j} = \sum_{i=1}^r n_{ij}, j = 1, 2, \dots, s \quad n_{\cdot \cdot} = \sum_{j=1}^s n_{\cdot j} = \sum_{i=1}^r n_{i\cdot} = \sum_{i=1}^r \sum_{j=1}^s n_{ij}$$

- (2) **χ^2 独立性检验** 假设检验问题 $H_0: p_{ij} = p_{i\cdot} \cdot p_{\cdot j}$ 。检验统计量

$$\chi^2 = \sum_{i,j} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} = \sum_{i,j} \frac{(n_{ij})^2}{e_{ij}} - n_{\cdot \cdot} \quad (4.4.2)$$

其中

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

统计量 $\chi^2 \rightarrow \chi^2((r-1)(s-1))$, 当 χ^2 取大值或 p 值很小的时候, 拒绝零假设。

第五章 回归分析

5.1 回归分析的基本概念

(1) 函数关系与相关关系

- (i) **函数关系** 变量之间确实存在的, 且在数量上表现为确定性的相互依存关系。
- (ii) **相关关系** 变量之间确实存在的, 但在数量上表现为不确定的相互依存关系。

$$y = f(x_1, x_2, \dots, x_k) + \epsilon \quad (5.1.1)$$

其中, x_1, x_2, \dots, x_k 称为解释变量 (自变量), y 称为被解释变量 (响应变量、因变量), ϵ 是其他随即影响因素, 通常假设是不可观测的随机误差, 是一个随机变量。

(2) 回归分析 回归分析就是研究相关关系的一种重要的数理统计方法。

(i) 分类

- 当自变量只有两个时, 称为一元回归分析; 当自变量在两个以上时, 称为多元回归分析。
- 变量间成线性关系, 称线性回归; 变量间不具有线性关系, 称非线性回归。

(ii) 多元线性回归模型

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \epsilon \quad (5.1.2)$$

通常假定 $\epsilon \sim N(0, \sigma^2)$ 。

5.2 一元回归

5.2.1 一元线性回归

(1) 一元线性回归模型

$$y = \beta_0 + \beta_1 x + \epsilon \quad (5.2.1)$$

(2) 一元线性回归方程

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \quad (5.2.2)$$

(3) 最小二乘估计 设 $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ 是 (x, y) 的一组观测值, 对每个样本观测值 (x_i, y_i) 考虑 y_i 与其回归值

$$E(y_i) = \beta_0 + \beta_1 x_i \quad (5.2.3)$$

的离差

$$y_i - E(y_i) = y_i - \beta_0 - \beta_1 x_i \quad (5.2.4)$$

定义离差平方和

$$Q(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - E(y_i))^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 \quad (5.2.5)$$

使其最小化。由

$$\begin{aligned}\left. \frac{\partial Q}{\partial \beta_0} \right|_{(\hat{\beta}_0, \hat{\beta}_1)} &= 0 \\ \left. \frac{\partial Q}{\partial \beta_1} \right|_{(\hat{\beta}_0, \hat{\beta}_1)} &= 0\end{aligned}$$

得正规方程组

$$\begin{cases} n\hat{\beta}_0 + n\bar{x}\hat{\beta}_1 = n\bar{y} \\ n\bar{x}\hat{\beta}_0 + \left(\sum_{i=1}^n x_i^2\right)\hat{\beta}_1 = \sum_{i=1}^n x_i y_i \end{cases} \quad (5.2.6)$$

解得

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (5.2.7)$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad (5.2.8)$$

(i) 简记

$$L_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2 \quad (5.2.9)$$

$$L_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y} \quad (5.2.10)$$

$$L_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - n\bar{y}^2 \quad (5.2.11)$$

则

$$\hat{\beta}_1 = \frac{L_{xy}}{L_{xx}} \quad (5.2.12)$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad (5.2.13)$$

(ii) 矩阵表示 记

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \quad A = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} \quad (5.2.14)$$

则一元回归方程的最小二乘解为

$$A = (X^T X)^{-1} X^T y \quad (5.2.15)$$

(iii) 最小二乘估计量的性质

- $\hat{\beta}_0, \hat{\beta}_1$ 都是 y_1, y_2, \dots, y_n 的线性组合;
- $\hat{\beta}_0, \hat{\beta}_1$ 的最小二乘估计都是无偏的;
- 有

$$\hat{\beta}_0 \sim N\left(\beta_0, \left(\frac{1}{n} + \frac{\bar{x}}{L_{xx}}\right) \sigma^2\right) \quad (5.2.16)$$

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{L_{xx}}\right) \quad (5.2.17)$$

$$\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\bar{x}}{L_{xx}} \sigma^2 \quad (5.2.18)$$

且 $\hat{\sigma}^2$ 与 $(\hat{\beta}_0, \hat{\beta}_1)$ 相互独立。

(4) σ^2 的估计 σ^2 的无偏估计为

$$\hat{\sigma}^2 = \frac{Q_e}{n-2} \quad (5.2.19)$$

其中 Q_e 使残差平方和，即因随机因素引起的误差

$$Q_e = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = L_{yy} - \hat{\beta}_1 L_{xy} \quad (5.2.20)$$

(5) 线性回归方程的显著性检验

(i) F 检验 总离差平方和、回归平方和、残差平方和满足 $S_{\text{总}}^2 = S_{\text{回}}^2 + S_{\text{残}}^2$ ，其中

$$S_{\text{总}}^2 = \sum_{i=1}^n (y_i - \bar{y})^2 \quad (5.2.21)$$

$$S_{\text{回}}^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \quad (5.2.22)$$

$$S_{\text{残}}^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (5.2.23)$$

- $S_{\text{回}}^2$ 反映了由于自变量 x 的变化引起的因变量 y 的差异，体现了 x 对 y 的影响；
- $S_{\text{残}}^2$ 反映了种种其它因素对 y 的影响，这些因素没有反映在自变量中，它们可作为随机因素看待。
- 当 $\beta_1 = 0$ 时， $S_{\text{回}}^2$ 与 $S_{\text{残}}^2$ 相互独立，且

$$\begin{aligned} \frac{S_{\text{回}}^2}{\sigma^2} &= \frac{1}{\sigma^2} \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \hat{\beta}_1^2 L_{xx} = \hat{\beta}_1 L_{xy} \\ &\sim \chi^2(1) \end{aligned}$$

$$\begin{aligned} \frac{S_{\text{残}}^2}{\sigma^2} &= \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = L_{yy} - \hat{\beta}_1 L_{xy} \\ &\sim \chi^2(n-2) \end{aligned}$$

检验假设 $H_0: \beta_1 = 0, H_1: \beta_1 \neq 0$ ，统计量

$$F = \frac{(n-1)S_{\text{回}}^2}{S_{\text{残}}^2} \sim F(1, n-2) \quad (5.2.24)$$

对于给定的显著性水平 $\alpha (0 < \alpha < 1)$ ， H_0 的拒绝域为

$$F > F_{\alpha}(1, n-2) \quad (5.2.25)$$

(ii) t 检验 检验假设 $H_0: \beta_1 = 0, H_1: \beta_1 \neq 0$ 。当原假设成立时，有 $\hat{\beta}_1 \sim N(0, \frac{\sigma^2}{L_{xx}})$ 即

$$\frac{\hat{\beta}_1 \sqrt{L_{xx}}}{\sigma} \sim N(0, 1)$$

又有

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

则有

$$t = \frac{\frac{\hat{\beta}_1 \sqrt{L_{xx}}}{\sigma}}{\sqrt{\frac{\frac{Q_e}{\sigma^2}}{(n-2)}}} = \frac{\hat{\beta}_1 \sqrt{L_{xx}}}{\hat{\sigma}} \sim t(n-2) \quad (5.2.26)$$

拒绝域为 $|t| \geq t_{\frac{\alpha}{2}}(n-2)$ 。

(iii) r 检验 相关系数

$$r^2 = \frac{L_{xy}^2}{L_{xx}L_{yy}} = \frac{S_{\text{回}}}{S_{\text{总}}} \quad (5.2.27)$$

r^2 越大, 说明 x 的变化引起 y 的变化就越大。

- $|r| \leq 1$;
- $|r| \neq 0$ 时, y 与 x 具有线性相关关系;
- $|r|$ 越大, 变量 y 与 x 之前的线性相关程度越强。

(6) 回归系数的区间估计

(i) $\hat{\beta}_1$ 置信水平为 $1 - \alpha$ 的置信区间为

$$\left(\hat{\beta}_1 - \frac{\hat{\sigma}}{\sqrt{L_{xx}}} t_{\frac{\alpha}{2}}(n-2), \hat{\beta}_1 + \frac{\hat{\sigma}}{\sqrt{L_{xx}}} t_{\frac{\alpha}{2}}(n-2) \right) \quad (5.2.28)$$

(ii) $\hat{\beta}_0$ 置信水平为 $1 - \alpha$ 的置信区间为

$$\left(\hat{\beta}_0 - \hat{\sigma} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{L_{xx}}} t_{\frac{\alpha}{2}}(n-2), \hat{\beta}_0 + \hat{\sigma} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{L_{xx}}} t_{\frac{\alpha}{2}}(n-2) \right) \quad (5.2.29)$$

(7) 预测

(i) 单值预测 对于任意给定的 $x = x_0$, y_0 的预测值为 $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$ 。

(ii) 区间预测 置信水平为 $1 - \alpha$ 的置信区间为

$$\left(\hat{y}_0 - \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{L_{xx}}} t_{\frac{\alpha}{2}}(n-2), \hat{y}_0 + \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{L_{xx}}} t_{\frac{\alpha}{2}}(n-2) \right) \quad (5.2.30)$$

当 n 较大时, $t_{\frac{\alpha}{2}}(n-2) \approx u_{\frac{\alpha}{2}}$, $1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{L_{xx}} \approx 1$, 则置信水平为 $1 - \alpha$ 的置信区间为

$$(\hat{y}_0 - \hat{\sigma} u_{\frac{\alpha}{2}}, \hat{y}_0 + \hat{\sigma} u_{\frac{\alpha}{2}}) \quad (5.2.31)$$

(8) 控制 要求 $y = \beta_0 + \beta_1 x + \epsilon$ 的值以 $1 - \alpha$ 的概率落在指定区间 (y', y'') 。控制 x 满足以下两个不等式

$$\hat{y} - \delta(x') \geq y' \quad (5.2.32)$$

$$\hat{y} + \delta(x'') \leq y'' \quad (5.2.33)$$

成立, 则区间 (x', x'') 就是所求 x 的控制区间。

5.2.2 非线性回归

表 5.1 正态总体均值和方差的假设检验

类型	表达式	变化	线性表达式
双曲线	$\frac{1}{y} = a + \frac{b}{x}$	$x' = \frac{1}{x}, y' = \frac{1}{y}$	$y' = a + bx'$

(续表)

类型	表达式	变化	线性表达式
幂函数	$y = ax^b$	$y' = \ln y, \beta_0 = \ln a, x' = \ln x, \beta_1 = b$ $\ln y = \ln a + b \ln x$	$y' = \beta_0 + \beta_1 x'$
指数曲线	$y = ae^{bx}$	$y' = \ln y, a' = \ln a$ $\ln y = \ln a + bx$	$y' = a' + bx$
倒指数曲线	$y = ae^{\frac{b}{x}}$	$y' = \ln y, a' = \ln a, x' = \frac{1}{x}$ $\ln y = \ln a + \frac{b}{x}$	$y' = a' + bx'$
对数曲线	$y = a + b \ln x$	$x' = \ln x$	$y = a + bx'$
S 型 (Logistic) 曲线	$y = \frac{K}{1 + Ae^{-\lambda x}}$	$y' = \ln \frac{K - y}{y}, a = \ln A$	$y' = a - \lambda x$

第六章 方差分析

6.1 单因素实验的方差分析

(1) 数学模型 设在实验中，因素 A 有 S 个不同水平

$$A_1, A_2, \dots, A_s$$

在水平下的实验结果 $X_j \sim N(\mu_j, \sigma^2) (j = 1, 2, \dots, s)$ 。其中 μ_j 和 σ^2 是位置参数。在水平 A_j 下做 n_j 次独立实验，得到结果

序号	A_1	A_2	\dots	A_s
1	X_{11}	X_{12}	\dots	X_{1s}
2	X_{21}	X_{22}	\dots	X_{2s}
3	X_{31}	X_{32}	\dots	X_{3s}
\vdots	\vdots	\vdots	\ddots	\vdots
n_j	$X_{n_1 1}$	$X_{n_2 2}$	\dots	$X_{n_s s}$
均值	$\bar{X}_{\bullet 1}$	$\bar{X}_{\bullet 2}$	\dots	$\bar{X}_{\bullet s}$

$X_{1j}, X_{2j}, \dots, X_{n_j j}$ 是来自总体 X_j 的容量为 n_j 的一个样本，其观察值为 $x_{1j}, x_{2j}, \dots, x_{n_j j}$ 。检验假设

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_s$$

$$H_1 : \mu_1, \mu_2, \dots, \mu_n \text{ 不全相等}$$

由于 X_{ij} 相互独立，且 $X_{ij} \sim N(\mu_j, \sigma^2) (i = 1, 2, \dots, n_j; j = 1, 2, \dots, s)$ ，若记 $\epsilon_{ij} = X_{ij} - \mu_j$ 则 $\epsilon_{ij} \sim N(0, \sigma^2)$ 且互相独立。则有

$$\begin{cases} X_{ij} = \mu_j + \epsilon_{ij} \\ \epsilon_{ij} \sim N(0, \sigma^2) \end{cases} \quad (6.1.1)$$

其中， $i = 1, 2, \dots, n_j$ ， $j = 1, 2, \dots, s$ ， μ_j 与 σ^2 均为未知参数。

(2) 线性可加模型 令

$$\begin{cases} \mu = \frac{1}{n} \sum_{j=1}^s n_j \mu_j, & i = 1, 2, \dots, m \\ \delta_j = \mu_j - \mu, & n = \sum_{j=1}^s n_j \end{cases}$$

则 μ 是各水平下总体均值的加权平均，称为总平均值； δ_i 代表了第 j 水平下的总体均值和平均值的差异，称为 A_j 的效应，满足

$$\sum_{j=1}^s n_j \delta_j = 0$$

得到

$$\begin{cases} X_{ij} = \mu + \delta_j + \epsilon_{ij} \\ \sum_{i=1}^m n_j \delta_j = 0 \end{cases} \quad (6.1.2)$$

其中, $i = 1, 2, \dots, n_j$, $j = 1, 2, \dots, s$, $\epsilon_{ij} \sim N(0, \sigma^2)$ 且相互独立。该模型的等价假设为

$$H_0: \delta_1, \delta_2, \dots, \delta_s = 0$$

$$H_1: \exists j, \delta_j \neq 0$$

(3) 总平方和的分解 总离差平方和为

$$\begin{aligned} S_T &= \sum_{j=1}^s \sum_{i=1}^{n_j} (X_{ij} - \bar{X})^2 \\ &= \sum_{j=1}^s \sum_{i=1}^{n_j} [(X_{ij} - \bar{X}_{\bullet j}) - (\bar{X}_{\bullet j} - \bar{X})]^2 \\ &= \sum_{j=1}^s \sum_{i=1}^{n_j} [(X_{ij} - \bar{X}_{\bullet j})]^2 - \sum_{j=1}^s \sum_{i=1}^{n_j} [(\bar{X}_{\bullet j} - \bar{X})]^2 \\ &= S_E + S_A \end{aligned} \quad (6.1.3)$$

(i) 组内离差和组间离差

- S_E 组内离差,

$$S_E = \sum_{j=1}^s \sum_{i=1}^{n_j} [(X_{ij} - \bar{X}_{\bullet j})]^2 \quad (6.1.4)$$

反映因素 A 各水平下的子样均值和样本值之间的差异, 是由随机误差引起的, 也叫误差平方和。

- S_A 组间离差,

$$S_A = \sum_{j=1}^s \sum_{i=1}^{n_j} [(\bar{X}_{\bullet j} - \bar{X})]^2 = \sum_{j=1}^s n_j (\bar{X}_{\bullet j} - \bar{X})^2 \quad (6.1.5)$$

反映因素 A 各水平下的子样均值和总平均值之间的差异, 也叫因素 A 效应的平方和。

(ii) S_T, S_E, S_A 的统计特性

$$\frac{S_T}{\sigma^2} \sim \chi^2(n-1) \quad (6.1.6)$$

$$\frac{S_E}{\sigma^2} \sim \chi^2(n-s) \quad (6.1.7)$$

$$\frac{S_A}{\sigma^2} \sim \chi^2(s-1) \quad (6.1.8)$$

(4) 检验方法 在 $H_0: \delta_i = 0$ 成立的条件下, 统计量

$$F = \frac{\frac{S_A}{\sigma^2} / (s-1)}{\frac{S_E}{\sigma^2} / (n-s)} \quad (6.1.9)$$

有 $F \sim F(s-1, n-s)$ 。给定 α 则:

- 若 $F > F_\alpha(s-1, n-s)$, 则拒绝 H_0 ;
- 若 $F < F_\alpha(s-1, n-s)$, 则接受 H_0 。

(5) 一元方差分析表 如表 6.1 所示

表 6.1 一元方差分析表

方差来源	平方和	自由度	均方	F 值
因素 A	S_A	$s - 1$	$\bar{S}_A = \frac{S_A}{s - 1}$	$F = \frac{\bar{S}_A}{\bar{S}_E}$
误差 E	S_E	$n - s$	$\bar{S}_E = \frac{S_E}{n - s}$	
总和 T	S_T	$n - 1$		

(6) 未知参数的估计 未知参数 μ, μ_j, σ^2

$$\hat{\sigma}^2 = \frac{S_E}{n - s} \quad (6.1.10)$$

$$\hat{\mu} = \bar{X} \quad (6.1.11)$$

$$\hat{\mu}_j = \bar{X}_{\bullet j} \quad (6.1.12)$$

6.2 双因素方差分析

6.2.1 非重复实验双因素方差分析

(1) 非重复实验双因素方差分析 有 A, B 两种因素, 因素 A 有 r 种水平, 因素 B 有 s 种水平, 在每种组合水平 $A_i \times B_j$ 上做了一次实验获得了实验值, 结果为 $X_{ij}, i = 1, 2, \dots, r, j = 1, 2, \dots, s$, X_{ij} 相互独立。

	B_1	B_2	\dots	B_n	$\bar{X}_{i\bullet}$
A_1	X_{11}	X_{12}	\dots	X_{1s}	$\bar{X}_{1\bullet}$
A_2	X_{21}	X_{22}	\dots	X_{2s}	$\bar{X}_{2\bullet}$
\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
A_r	X_{r1}	X_{r2}	\dots	X_{rs}	$\bar{X}_{r\bullet}$

(2) 数学模型 假设总体服从正态分布 $N(\mu_{ij}, \sigma^2)$, 令 $X_{ij} = \mu_{ij} + \sigma_{ij}$ 则 $\epsilon_{ij} \sim N(0, \sigma^2)$, ϵ_{ij} 相互独立。令

$$\mu = \frac{1}{rs} \sum_{i=1}^r \sum_{j=1}^s \mu_{ij}$$

$$\mu_{i\bullet} = \frac{1}{s} \sum_{j=1}^s \mu_{ij}$$

$$\mu_{\bullet j} = \frac{1}{r} \sum_{i=1}^r \mu_{ij}$$

$$\alpha_i = \mu_{i\bullet} - \mu$$

$$\beta_j = \mu_{\bullet j} - \mu$$

$$\gamma_{ij} = (\mu_{ij} - \mu) - \alpha_i - \beta_j = \mu_{ij} - \mu_{i\bullet} - \mu_{\bullet j} + \mu$$

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}, i = 1, 2, \dots, r; j = 1, 2, \dots, s$$

$$X_{ij} = \mu_{ij} + \epsilon_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ij}, \gamma_{ij} = 0$$

α_i 称为因素 A 在水平 A_i 的效应, β_i 称为因素 B 在水平 B_i 的效应。检验假设:

$$H_{01} : \alpha_1 = \alpha_2 = \cdots = \alpha_n = 0$$

$$H_{02} : \beta_1 = \beta_2 = \cdots = \beta_n = 0$$

(3) 平方和分解

$$\begin{aligned} S_T &= \sum_{i=1}^r \sum_{j=1}^s (X_{ij} - \bar{X})^2 \\ &= s \sum_{i=1}^r (\bar{X}_{i\bullet} - \bar{X})^2 + r \sum_{j=1}^s (\bar{X}_{\bullet j} - \bar{X})^2 + \sum_{i=1}^r \sum_{j=1}^s (X_{ij} - \bar{X}_{i\bullet} - \bar{X}_{\bullet j} + \bar{X})^2 \\ &= S_A + S_B + S_E \end{aligned} \quad (6.2.1)$$

当 H_{01}, H_{02} 成立时,

$$\frac{S_A}{\sigma^2} \sim \chi^2(r-1) \quad (6.2.2)$$

$$\frac{S_B}{\sigma^2} \sim \chi^2(s-1) \quad (6.2.3)$$

$$\frac{S_T}{\sigma^2} \sim \chi^2(rs-1) \quad (6.2.4)$$

$$\frac{S_E}{\sigma^2} \sim \chi^2(rs-r-s+1) \quad (6.2.5)$$

(4) 均方离差

- 因素 A 引起的均方离差

$$\bar{S}_A = \frac{S_A}{r-1}$$

- 因素 B 引起的均方离差

$$\bar{S}_B = \frac{S_B}{s-1}$$

- 均方误差

$$\bar{S}_E = \frac{S_E}{(r-1)(s-1)}$$

(5) 拒绝域 给定显著性水平 α , 检验统计量和拒绝域

$$H_{00} : F_A = \frac{\bar{S}_A}{\bar{S}_E} \sim F(r-1, (r-1)(s-1)) \quad F_A \geq F_\alpha(r-1, (r-1)(s-1)) \quad (6.2.6)$$

$$H_{01} : F_B = \frac{\bar{S}_B}{\bar{S}_E} \sim F(r-1, (r-1)(s-1)) \quad F_B \geq F_\alpha(s-1, (r-1)(s-1)) \quad (6.2.7)$$

(6) 方差分析表 如表 6.2 所示。

6.2.1.1 重复实验方差双因素方差分析

(1) 重复实验方差双因素方差分析 有 A, B 两种因素, 因素 A 有 r 种水平, 因素 B 有 s 种水平, 在每种组合水平 $A_i \times B_j$ 上进行 c 次实验, 结果为 $X_{ijk} (i=1, 2, \cdots, r; j=1, 2, \cdots, s; k=1, 2, \cdots, c)$, X_{ijk} 相互

表 6.2 非重复实验双因素方差分析表

方差来源	平方和	自由度	均方	F 值
因素 A	S_A	$r - 1$	$\bar{S}_A = \frac{S_A}{r - 1}$	$F_A = \frac{\bar{S}_A}{\bar{S}_E}$
因素 B	S_B	$s - 1$	$\bar{S}_B = \frac{S_B}{s - 1}$	$F_B = \frac{\bar{S}_B}{\bar{S}_E}$
误差 E	S_E	$(r - 1)(s - 1)$	$\bar{S}_E = \frac{S_E}{n - s}$	
总和 T	S_T	$rs - 1$		

独立。

	B_1	B_2	\cdots	B_n	$\bar{X}_{i\bullet\bullet}$
A_1	$X_{111}, X_{112}, \cdots, X_{11c}$	$X_{121}, X_{122}, \cdots, X_{12c}$	\cdots	$X_{1s1}, X_{1s2}, \cdots, X_{1sc}$	$\bar{X}_{1\bullet\bullet}$
A_2	$X_{211}, X_{212}, \cdots, X_{21c}$	$X_{221}, X_{222}, \cdots, X_{22c}$	\cdots	$X_{2s1}, X_{2s2}, \cdots, X_{2sc}$	$\bar{X}_{2\bullet\bullet}$
\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
A_r	$X_{r11}, X_{r12}, \cdots, X_{r1c}$	$X_{r21}, X_{r22}, \cdots, X_{r2c}$	\cdots	$X_{rs1}, X_{rs2}, \cdots, X_{rsc}$	$\bar{X}_{r\bullet\bullet}$
$\bar{X}_{\bullet j\bullet}$	$\bar{X}_{\bullet 1\bullet}$	$\bar{X}_{\bullet 2\bullet}$	\cdots	$\bar{X}_{\bullet s\bullet}$	\bar{X}

(2) 数学模型 假设总体服从正态分布 $X_{ijk} \sim N(\mu_{ij}, \sigma^2)$ ，令 $X_{ijk} = \mu_{ij} + \epsilon_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$ ，则 $\epsilon_{ijk} \sim N(0, \sigma^2)$ ， ϵ_{ijk} 相互独立。令

$$\begin{aligned}\mu &= \frac{1}{rs} \sum_{i=1}^r \sum_{j=1}^s \mu_{ij} \\ \mu_{i\bullet} &= \frac{1}{s} \sum_{j=1}^s \mu_{ij} \\ \mu_{\bullet j} &= \frac{1}{r} \sum_{i=1}^r \mu_{ij} \\ \alpha_i &= \mu_{i\bullet} - \mu \\ \beta_j &= \mu_{\bullet j} - \mu \\ \gamma_{ij} &= (\mu_{ij} - \mu) - \alpha_i - \beta_j = \mu_{ij} - \mu_{i\bullet} - \mu_{\bullet j} + \mu \\ \mu_{ij} &= \mu + \alpha_i + \beta_j + \gamma_{ij}, i = 1, 2, \cdots, r; j = 1, 2, \cdots, s \\ X_{ij} &= \mu_{ij} + \epsilon_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ij}, \gamma_{ij} = 0\end{aligned}$$

α_i 称为因素 A 在水平 A_i 的效应， β_j 称为因素 B 在水平 B_j 的效应， γ_{ij} 称为因素 A, B 在组合水平 $A_i \times B_j$ 的交互效应。检验假设：

$$\begin{aligned}H_{01} &: \alpha_1 = \alpha_2 = \cdots = \alpha_n = 0 \\ H_{02} &: \beta_1 = \beta_2 = \cdots = \beta_n = 0 \\ H_{03} &: \gamma_{ij} = 0, i = 1, 2, \cdots, r; j = 1, 2, \cdots, s\end{aligned}$$

(3) 平均数

$$\bar{X}_{ij\bullet} = \frac{1}{c} \sum_{k=1}^c X_{ijk} \quad (6.2.8)$$

$$\bar{X}_{i\bullet\bullet} = \frac{1}{sc} \sum_{j=1}^s \sum_{k=1}^c X_{ijk} \quad (6.2.9)$$

$$\bar{X}_{\bullet j\bullet} = \frac{1}{rc} \sum_{i=1}^r \sum_{k=1}^c X_{ijk} \quad (6.2.10)$$

$$\bar{X} = \frac{1}{rsc} \sum_{i=1}^r \sum_{j=1}^s \sum_{k=1}^c X_{ijk} \quad (6.2.11)$$

(4) 平方和分解

$$\begin{aligned} S_T &= \sum_{i=1}^r \sum_{j=1}^s \sum_{k=1}^c (X_{ijk} - \bar{X})^2 \\ &= sc \sum_{i=1}^r (\bar{X}_{i\bullet\bullet} - \bar{X})^2 + rc \sum_{j=1}^s (\bar{X}_{\bullet j\bullet} - \bar{X})^2 + c \sum_{i=1}^r \sum_{j=1}^s (X_{ij\bullet} - \bar{X}_{i\bullet\bullet} - \bar{X}_{\bullet j\bullet} + \bar{X})^2 + \sum_{i=1}^r \sum_{j=1}^s \sum_{k=1}^c (X_{ijk} - \bar{X}_{ij\bullet})^2 \\ &= S_A + S_B + S_I + S_E \end{aligned} \quad (6.2.12)$$

- 因素 A 的离差平方和

$$S_A = sc \sum_{i=1}^r (\bar{X}_{i\bullet\bullet} - \bar{X})^2$$

- 因素 B 的离差平方和

$$S_B = rc \sum_{j=1}^s (\bar{X}_{\bullet j\bullet} - \bar{X})^2$$

- 因素 A, B 交互作用引起的离差平方和

$$S_I = c \sum_{i=1}^r \sum_{j=1}^s (X_{ij\bullet} - \bar{X}_{i\bullet\bullet} - \bar{X}_{\bullet j\bullet} + \bar{X})^2$$

- 误差平方和

$$S_E = \sum_{i=1}^r \sum_{j=1}^s \sum_{k=1}^c (X_{ijk} - \bar{X}_{ij\bullet})^2$$

(5) 离差平方和的统计性质 当 H_{01}, H_{02}, H_{03} 成立时,

$$\frac{S_A}{\sigma^2} \sim \chi^2(r-1) \quad (6.2.13)$$

$$\frac{S_B}{\sigma^2} \sim \chi^2(s-1) \quad (6.2.14)$$

$$\frac{S_I}{\sigma^2} \sim \chi^2((r-1)(s-1)) \quad (6.2.15)$$

$$\frac{S_E}{\sigma^2} \sim \chi^2(rs(c-1)) \quad (6.2.16)$$

$$\frac{S_T}{\sigma^2} \sim \chi^2(rsc-1) \quad (6.2.17)$$

(6) 均方离差

$$\bar{S}_A = \frac{S_A}{r-1} \quad (6.2.18)$$

$$\bar{S}_B = \frac{S_B}{s-1} \quad (6.2.19)$$

$$\bar{S}_I = \frac{S_I}{(r-1)(s-1)} \quad (6.2.20)$$

$$\bar{S}_E = \frac{S_E}{rs(c-1)} \quad (6.2.21)$$

(7) 检验统计量和拒绝域

$$H_{01} : F_A = \frac{\bar{S}_A}{\bar{S}_E} \sim F(r-1, rs(c-1)) \quad F_A \geq F_\alpha(r-1, rs(c-1)) \quad (6.2.22)$$

$$H_{02} : F_B = \frac{\bar{S}_B}{\bar{S}_E} \sim F(s-1, rs(c-1)) \quad F_A \geq F_\alpha(s-1, rs(c-1)) \quad (6.2.23)$$

$$H_{03} : F_I = \frac{\bar{S}_I}{\bar{S}_E} \sim F((r-1)(s-1), rs(c-1)) \quad F_A \geq F_\alpha((r-1)(s-1), rs(c-1)) \quad (6.2.24)$$

(8) 方差检验表 如表 所示。

表 6.3 重复实验双因素方差分析表

方差来源	平方和	自由度	均方	F 值
因素 A	S_A	$r-1$	$\bar{S}_A = \frac{S_A}{r-1}$	$F_A = \frac{\bar{S}_A}{\bar{S}_E}$
因素 B	S_B	$s-1$	$\bar{S}_B = \frac{S_B}{s-1}$	$F_B = \frac{\bar{S}_B}{\bar{S}_E}$
交互作用 I	S_I	$(r-1)(s-1)$	$\bar{S}_I = \frac{S_I}{(r-1)(s-1)}$	$F_I = \frac{\bar{S}_I}{\bar{S}_E}$
误差 E	S_E	$rs(c-1)$	$\bar{S}_E = \frac{S_E}{rs(c-1)}$	
总和 T	S_T	$rsc-1$		