概率论与数理统计教程

经典书籍重排

作者：茆诗松　程依名　濮晓龙

组织：RE-Book Program

时间：April 16, 2019

版本：3.00

普通高等教育“十五”国家级规划教材

目　　录

第 1 章 随机事件与概率

随机事件

第 2 章 随机变量及其分布

随机变量及其分布

第 3 章 多维随机变量及其分布

在有些随机现象中, 对每个样本点 ω 只用一个随机变量去描述是不够的, 譬如要研究儿童的生长发育情况, 仅研究儿童的身高 $X(\omega)$ 或仅研究其体重 $Y(\omega)$ 都是片面的, 有必要把 $X(\omega)$ 和 $Y(\omega)$ 作为一个整体来考虑, 讨论它们总体变化的统计规律性, 进一步可以讨论 $X(\omega)$ 与 $Y(\omega)$ 之间的关系, 在有些随机现象中, 甚至要同时研究二个以上随机变量.

如何来研究多维随机变量的统计规律性呢, 仿一维随机变量, 我们先研究联合分布函数, 然后研究离散随机变量的联合分布列、连续随机变量的联合密度函数.

3.1 多维随机变量及其联合分布

3.1.1 多维随机变量

下面我们先给出 n 维随机变量的定义.

定义 3.1.1 (随机变量). 如果 $X_1(\omega), X_2(\omega), \ldots, X_n(\omega)$ 是定义在同一样本空间 $\Omega = \{\omega\}$ 上的 n 个随机变量, 则称

$$X(\omega) = (X_1(\omega), X_2(\omega), \ldots, X_n(\omega))$$

为 n 维(或 n 元)随机变量或随机向量.

注意, 多维随机变量的关键是定义在同一样本空间上, 对于不同样本空间上的两个随机变量, 我们只能在乘积空间 $\Omega_1 \times \Omega_2 = \{(\omega_1, \omega_2); \omega_1 \in \Omega_1, \omega_2 \in \Omega_2\}$ 上讨论, 这要用到更多的工具, 本章将不涉及这类问题.

在实际问题中, 多维随机变量的情况是经常会遇到的譬如

- 在研究四岁至六岁儿童的生长发育情况时, 我们感兴趣于每个儿童(样本点 ω)的身高 $X_1(\omega)$ 和体重 $X_2(\omega)$. 这里 $(X_1(\omega), X_2(\omega))$ 是一个二维随机变量.
- 在研究每个家庭的支出情况时, 我们感兴趣于每个家庭(样本点 ω)的衣食住行四个方面, 若用 $X_1(\omega), X_2(\omega), X_3(\omega), X_4(\omega)$ 分别表示衣食住行的花费占其家庭总收人的百分比, 则 $X_1(\omega), X_2(\omega), X_3(\omega), X_4(\omega)$ 就是一个四维随机变量.

3.1.2 联合分布函数

定义 3.1.2 (联合分布函数). 对任意的 n 个实数 $x_1, x_2, \ldots, x_n$, 则 n 个事件 $\{X_1 \le x_1\}, \{X_2 \le x_2\}, \ldots, \{X_n \le x_n\}$ 同时发生的概率

$$F(x_1, x_2, \cdots, x_n) = P(X_1 \le x_1, X_2 \le x_2, \ldots, X_n \le x_n) \tag{3.1.1}$$

称为 n 维随机变量 $(X_1, X_2, \ldots, X_n)$ 的联合分布函数.

本章主要研究二维随机变量, 二维以上的情况可类似进行.

在二维随机变量 (X, Y) 场合, 联合分布函数 $F(x, y) = P(X \le x, Y \le y)$ 是事件 $\{X \le x\}$ 与 $\{Y \le y\}$ 同时发生(交)的概率. 如果将二维随机变量 (X, Y) 看成是平面上随机点的坐标, 那么联合分布函数 $F(x, y)$ 在 (x, y) 处的函数值就是随机点 (X, Y) 落在以 (x, y) 为右上角的无穷矩形内的概率, 见图 **??** .

定理 3.1.1. 任一二维联合分布函数 $F(x, y)$ 必具有如下四条基本性质:

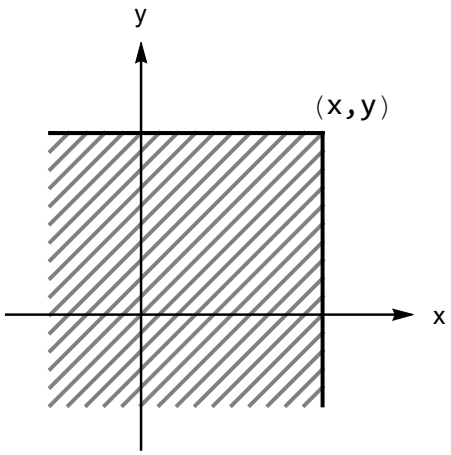

图 3.1.1: 联合分布函数示意图

1. 单调性　$F(x,y)$ 分别对 x 或 y 是单调不减的，即
   - 当 $x_1 < x_2$ 时，有 $F(x_1,y) \le F(x_2,y)$.
   - 当 $y_1 < y_2$ 时，有 $F(x,y_1) \le F(x,y_2)$.
2. 有界性　对任意的 x 和 y，有 $0 \le F(x,y) \le 1$，且

$$F(-\infty,y) = \lim_{x\to-\infty} F(x,y) = 0,$$
$$F(x,-\infty) = \lim_{y\to-\infty} F(x,y) = 0,$$
$$F(+\infty,+\infty) = \lim_{x,y\to+\infty} (x,y) = 1.$$

3. 右连续性　对每个变量都是右连续的，即

$$F(x+0,y) = F(x,y),$$
$$F(x,y+0) = F(x,y).$$

4. 非负性　对任意的 $a < b, c < d$ 有

$$P(a < X \le b, c < Y \le d) = F(b,d) - F(a,d) - F(b,c) + F(a,c) \ge 0.$$

证明：

(a). 因为当 $x_1 < x_2$ 时，有 $\{X_1 \le x_1\} \subset \{X_2 \le x_2\}$，所以对任意给定的 y 有

$$\{X \le x_1, Y \le y\} \subseteq \{X \le x_2, Y \le y\},$$

由此可得

$$F(x_1,y) = P(X \le x_1, Y \le y) \le P(X \le x_2, Y \le y) = F(x_2,y),$$

即 $F(x,y)$ 关于 x 是单调不减的，同理可证 $F(x,y)$ 关于 y 是单调不减的.

(b). 由概率的性质可知 $0 \le F(x,y) \le 1$. 又因为对任意的正整数 n 有

$$\lim_{x\to-\infty} \{X \le x\} = \lim_{n\to+\infty} \bigcap_{m=1}^{n} \{X \le -m\} = \varnothing,$$
$$\lim_{x\to+\infty} \{X \le x\} = \lim_{n\to+\infty} \bigcup_{m=1}^{n} \{X \le m\} = \Omega,$$

对 $Y \le y$ 也类似可得. 再由概率的连续性, 就可得

$$F(-\infty, y) = F(x, -\infty) = 0; \quad F(+\infty, +\infty) = 1.$$

(c). 固定 y, 仿一维分布函数右连续的证明, 就可得知 $F(x,y)$ 关于 x 是右连续的. 同样固定 x 可证得 $F(x,y)$ 关于 y 是右连续的.

(d). 只需证

$$P(a < X \le b, c < Y \le d) = F(b,d) - F(a,d) - F(b,c) + F(a,c).$$

为此记 (见图 **??**)

$$A = \{X \le a\}, \quad B = \{X \le b\}, \quad C = \{Y \le c\}, \quad D = \{Y \le d\},$$

考虑到

$$\{a < X \le b\} = B - A = B \cap \overline{A}, \quad \{c < Y \le d\} = D - C = D \cap \overline{C},$$

且 $A \subset B, C \subset D$, 由此可得

$$\begin{aligned}
0 &\le P(a < X \le b, c < Y \le d) \\
&= P(B \cap \overline{A} \cap D \cap \overline{C}) \\
&= P(BD - (A \cup C)) \\
&= P(BD) - P(ABD \cup BCD) \\
&= P(BD) - P(AD \cup BC) \\
&= P(BD) - P(AD) - P(BC) + P(ABCD) \\
&= P(BD) - P(AD) - P(BC) + P(AC) \\
&= P(BD) - P(AD) - P(BC) + P(ABCD) \\
&= F(b,d) - F(a,d) - F(b,c) + F(a,c).
\end{aligned}$$

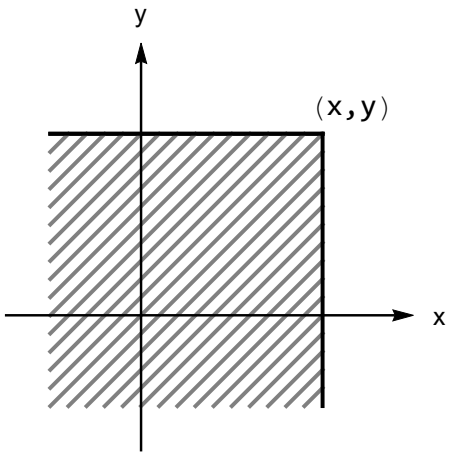

图 3.1.2: 二维随机变量 (X,Y) 落在矩形中的情况

第 4 章 大数定律与中心极限定理

大数定律与中心极限定理

4.1 特征函数

设 $p(x)$ 是随机变量 X 的密度函数, 则 $p(x)$ 的傅里叶变换是

$$\varphi(t) = \int_{-\infty}^{+\infty} \mathrm{e}^{itx} p(x)\,dx,$$

其中 $i = \sqrt{-1}$ 是虚数单位. 由数学期望的概念知, $\varphi(t)$ 恰好是 $E\left(\mathrm{e}^{itx}\right)$. 这就是本节要讨论的特征函数, 它是处理许多概率论问题的有力工具, 它能把寻求独立随机变量和的分布的卷积运算 (积分运算) 转换成乘法运算, 还能把求分布的各阶原点矩 (积分运算) 转换成微分运算. 特别它能把寻求随机变量序列的极限分布转换成一般的函数极限问题, 下面从特征函数的定义开始介绍它们.

4.1.1 特征函数的定义

定义 4.1.1. 设 X 是一个随机变量, 称

$$\varphi(t) = E\left(\mathrm{e}^{itx}\right),\ -\infty \le t \le +\infty, \tag{4.1.1}$$

为 X 的特征函数.

因为 $|\mathrm{e}^{itx}| \le 1$, 所以 $E\left(\mathrm{e}^{itX}\right)$ 总是存在的, 即任一随机变量的特征函数总是存在的.

当离散随机变量 X 的分布列为 $p_k = P(X = x_k)$, $k = 1, 2, \ldots$, 则 X 的特征函数为

$$\varphi(t) = \sum_{k=1}^{+\infty} \mathrm{e}^{itx_k} p_k,\ -\infty \le t \le +\infty. \tag{4.1.2}$$

当连续随机变量 X 的密度函数为 $p(x)$, 则 X 的特征函数为

$$\varphi(t) = \int_{-\infty}^{+\infty} \mathrm{e}^{itx} p(x)\,dx,\ -\infty \le t \le +\infty. \tag{4.1.3}$$

与随机变量的数学期望、方差及各阶矩一样, 特征函数只依赖于随机变量的分布, 分布相同则特征函数也相同, 所以我们也常称为某**分布的特征函数**.

例 4.1.1: 常用分布的特征函数

1. **单点分布**: $P(X = a) = 1$, 其特征函数为

$$\varphi(t) = \mathrm{e}^{itz}.$$

2. **0-1 分布**: $P(X = x) = p^x(1-p)^{1-x}$, $x = 0, 1$, 其特征函数为

$$\varphi(t) = p\mathrm{e}^{it} + q, \quad \text{其中 } q = 1 - p.$$

3. **泊松分布**: $P(X = k) = \frac{\lambda^k}{k!}\mathrm{e}^{-\lambda}$, $k = 0, 1, \ldots$, 其特征函数为

$$\varphi(t) = \sum_{k=0}^{+\infty} \mathrm{e}^{ikt}\frac{\lambda^k}{k!}\mathrm{e}^{-\lambda} = \mathrm{e}^{\lambda}\mathrm{e}^{\lambda\mathrm{e}^{it}} = \mathrm{e}^{\lambda(\mathrm{e}^{it})-1}.$$

4. **均匀分布** $U(a,b)$: 因为密度函数为

$$p(x)=\begin{cases}\dfrac{1}{b-a}, & a<x<b,\\ 0, & \text{其他}.\end{cases}$$

所以特征函数为

$$\varphi(t)=\int_a^b \frac{\mathrm{e}^{itx}}{b-a}\,dx=\frac{\mathrm{e}^{ibt}-\mathrm{e}^{iat}}{it(b-a)}.$$

5. **标准正态分布** $N(0,1)$: 因为密度函数为

$$p(x)=\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{x^2}{2}},\quad -\infty<x<+\infty.$$

所以特征函数为

$$\begin{aligned}\varphi(t)&=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\mathrm{e}^{itx-\frac{x^2}{2}}\,dx\\&=\mathrm{e}^{-\frac{t^2}{2}}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\mathrm{e}^{-\frac{(x-it)^2}{2}}\,dx\\&=\mathrm{e}^{-\frac{t^2}{2}}\frac{1}{\sqrt{2\pi}}\int_{-\infty-it}^{+\infty-it}\mathrm{e}^{-\frac{x^2}{2}}\,dx\\&=\mathrm{e}^{-\frac{t^2}{2}},\end{aligned}$$

其中

$$\int_{-\infty-it}^{+\infty-it}\mathrm{e}^{-\frac{x^2}{2}}\,dx=\sqrt{2\pi}$$

是利用复变函数中的围道积分求得的. 有了标准正态分布的特征函数, 再利用下节给出的特征函数的性质, 就很容易得到一般正态分布 $N(\mu,\sigma^2)$ 的特征函数, 见例 .

6. **指数分布** $\exp(\lambda)$: 因为密度函数为

$$p(x)=\begin{cases}\lambda\mathrm{e}^{-\lambda x}, & x>0,\\ 0, & x\le 0.\end{cases}$$

所以特征函数为

$$\begin{aligned}\varphi(t)&=\int_0^{+\infty}\mathrm{e}^{itx}\lambda\mathrm{e}^{-\lambda x}\,dx\\&=\lambda\left(\int_0^{+\infty}\cos(tx)\mathrm{e}^{-\lambda x}\,dx+i\int_0^{+\infty}\sin(tx)\mathrm{e}^{-\lambda x}\,dx\right)\\&=\lambda\left(\frac{\lambda}{\lambda^2+t^2}+i\frac{t}{\lambda^2+t^2}\right)\\&=\left(1-\frac{it}{\lambda}\right)^{-1}.\end{aligned}$$

以上积分中用到了复变函数中的欧拉公式: $\mathrm{e}^{itx}=\cos(tx)+i\sin(tx)$.

第 5 章 统计量及其分布

前四章的研究属于概率论的范畴. 我们已经看到, 随机变量及其概率分布全面地描述了随机现象的通解规律性. 在概率论的许多问题中, 概率分布通常被假定为已知的, 而一切计算及推理均基于这个已知的分布进行, 在实际问题中, 情况往往并非如此, 看一个例子.

例 5.0.1: 某公司要采购一批产品, 每件产品不是合格品就是不合格品, 但该产品总有一个不合格品率 p. 由此, 若从该批产品中随机抽取一件, 用 X 表示这一件产品的不合格数, 不难看出 X 服从一个二点分布 $b(1,p)$, 但分布中的参数 p 却是不知道的. 显然, p 的大小决定了该批产品的质量, 它直接影响采购行为的经济效益. 因此, 人们会对 p 提出一些问题, 比如,

- p 的大小如何;
- p 大概落在什么范围内;
- 能否认为 p 满足设定要求(如 $p \leq 0.05$).

诸如例 **??** 研究的问题属于数理统计的范畴. 接下来我们从统计中最基本的概念——总体和样本开始介绍统计学内容.

5.1 总体与样本

5.1.1 总体与个体

在一个统计问题中, 我们把研究对象的全体称为**总体**, 构成总体的每个成员称为**个体**. 对多数实际问题, 总体中的个体是一些实在的人或物. 比如, 我们要研究某大学的学生身高情况, 则该大学的全体学生构成问题的总体, 而每一个学生即使一个个体. 事实上, 每个学生都有许多特征: 性别、年龄、身高、体重、名字、籍贯等等, 而在该问题中, 我们关心的只是该校学生的身高如何, 对其他的特征暂不予考虑. 这样, 每个学生(个体)所具有的数量指标值——身高就是个体, 而将所有身高全体看成总体. 这样一来, 若抛开实际背景, 总体就是一堆数, 这堆数中有大有小, 有的出现的机会多, 有的出现机会小, 因此用一个概率分布去描述和归纳总体是恰当的, 从这个意义看, **总体就是一个分布**, 而其数量指标就是服从这个分布的随机变量. 以后说 ‘‘从总体中抽样" 与 ‘‘从分布中抽样" 是同一个意思.

例 5.1.2: 考察某厂的产品质量, 将其产品只分为合格品与不合格品, 并以 0 记合格品, 以 1 记不合格品, 则

$$\text{总体} = \{\text{该厂生产的全部合格品与不合格品}\} = \{\text{由 0 或 1 组成的一堆数}\}.$$

若以 p 表示这堆数中 1 的比例(不合格品率), 则该总体可由一个二点分布表示:

X	0	1
P	$1-p$	p

不同的 p 反映了总体间的差异. 譬如, 两个生产同类产品的工厂的产品总体分布为

X	0	1
P	0.983	0.017

X	0	1
P	0.915	0.085

我们可以看到, 第一个工厂的产品质量优于第二个工厂. 实际中, 分布中的不合格率是未知的, 如何对之进行估计是统计学要研究的问题.

例 5.1.3: 彩电的彩色浓度是彩电质量好坏的一个重要指标. 20 世纪 70 年代在美国销售的 SONY 牌彩电有两个产地: 美国和日本, 两地的工厂是按统一设计方案和相同的生产线生产同一型号 SONY 彩电, 连使用说明书和检验合格的标准也是一样的. 其中关于彩色浓度 X 的标准是: 目标值为 m, 公差为 5, 即当 X 在 $[m-5, m+5]$ 内该彩电的热情高于购买美产 SONY 彩电, 原因何在? 这就要考察这两个总体有什么差别. 1979 年 4 月 17 日日本《朝日新闻》刊登调查报告指出, 日产 SONY 彩电的彩色浓度服从正态分布 $N(m, (5/3)^2)$, 而美产 SONY 彩电的彩色浓度服从 $(m-5, m+5)$ 上的均匀分布, 见图 **??**. 这两个不同的分布代表了不同的总体, 其均值相同 (都为 m), 但方差不同. 若彩色浓度与 m 的距离在 5/3 以内为 I 级品, 在 5/3 到 10/3 之间为 II 级品, 在 10/3 到 5 之间为 III 级品, 其他为 IV 级品. 于是日产 SONY 彩电的 I 级品为美产 SONY 的两倍出头 (见表 **??**), 这就是美国消费者愿意购买日产 SONY 的主要原因.

在有些问题中, 我们对每一研究对象可能要观测两个甚至更多个指标, 此时可用多维随机向量及其联合分布来描述总体. 这种总体称为多维总体. 譬如, 我们要了解某校大学生的三个指标: 年龄、身高、月生活支出. 则我们可用一个三维随机向量描述该总体. 这是一个三维总体, 它是多元分析所研究的对象. 本书中主要研究一维总体, 某些地方也会涉及二维总体.

总体还有有限总体和无限总体, 本书将以无限总体作为主要研究对象.

5.1.2 样本

为了了解总体的分布, 我们从总体中随机地抽取 n 个个体, 记其指标值为 $x_1, x_2, \cdots, x_n$, 则 $x_1, x_2, \cdots, x_n$ 称为总体的一个**样本**, n 称为**样本容量**, 或简称为**样本量**, 样本中的个体称为**样品**.

第 6 章 参数估计

参数估计

第 7 章 假设检验

统计推断的另一个主要内容是统计假设检验。在这一章里我们将讨论统计假设的设立及其检验问题。

7.1 假设检验问题

我们从一个例子开始引出假设检验问题。

例题 7.1.1: 某厂生产的合金强度服从正态分布 $N(\theta,16)$，其中 θ 的设计值为不低于 110Pa . 为保证质量，该厂每天都要对生产情况做例行检查，以判断生产是否正常进行，即该合金的平均强度不低于 110(pa) . 某天从生产中随机抽取 25 块合金，测得强度值为 $x_1,x_2,\cdots,x_{25}$ 其均值为 $\bar{x}=108$(Pa)，问当日生产是否正常？

对这个实际问题可作如下分析；

(1) 这不是一个参数估计问题.

(2) 这是在给定总体与样本下，要求对命题“合金平均强度不低于 110Pa”作出回答:“是”还是“否”? 这类问题称为统计假设检验问题，简称假设检验问题.

(3) 命题:“合金平均强度不低于 110Pa”正确与否仅涉及参数 0，因此该命题是否正确将涉及如下两个参数集合：

$$\theta_0=|\theta;\theta\geq 110|,\theta_1=|\theta:\theta<110|$$

命题成立对应于 "$\theta\in\theta_0$"，命题不成立则对应 "$\theta\in\theta_1$" . 在统计学中这两个非空参数集合都称作统计假设，简称假设.

(4) 我们的任务是利用所给总体 $N(\theta,16)$ 和样本均值 $\bar{x}=108$(Pa) 去判断假设（命题 "$\theta\in\theta_0$" . 是否成立，这里的“判断”在统计学中称为检验或检验法则。

检验结果有两种：

“假设不正确”——称为拒绝该假设；

“假设正确”——称为接收该假设.

(5) 若假设可用一个参数的集合表示，该假设检验问题称为参数假设检验问题，否则称为非参数假设检验问题，例 7.1.1 就是一个参数假设检验问题，而对假设“总体为正态分布”作出检验的问题就是一个非参数假设检验问题. 本章前三节讲述参数假设检验问题，最后一节 (7.4) 将讨论非参数假设检验问题.

第 8 章 方差分析与回归分析

8.1 方差分析

8.1.1 问题的提出

前面几章我们讨论的都是一个总体或者两个总体的统计分析问题, 在实际工作中我们还会经常碰到多个总体均值的比较问题, 处理这类问题通常采用所谓的方差分析方法. 本节将叙述这个方法, 先看一个例子.

例 8.1.1: 在饲料养鸡增肥的研究中, 某研究所提出三种饲料配方: A_1 是以鱼粉为主的饲料, A_2 是以槐树粉为主的饲料, A_3 是以苜蓿粉为主的饲料. 为比较三种饲料的效果, 特选 24 只相似的雏鸡随机均分为三组, 每组各喂一种饲料, 60 天后观察它们的重量. 试验结果如下表所示:

表 8.1.1: 鸡饲料试验数据

饲料 A	鸡重/g							
A_1	1073	1009	1060	1001	1002	1012	1009	1028
A_2	1107	1092	990	1109	1090	1074	1122	1001
A_3	1093	1029	1080	1021	1022	1032	1029	1048

本例中, 我们要比较的是三种饲料对鸡的增肥作用是否相同. 为此, 把饲料称为**因子**, 记为 A, 三种不同的配方称为因子 A 的三个水平, 记为 A_1, A_2, A_3, 使用配方 A_i 下第 j 只鸡 60 天后的重量用 y_{ij} 表示, $i = 1,2,3$, $j = 1,2,3,\ldots,10$. 我们的目的是比较三种不同饲料配方下鸡的平均重量是否相等, 为此, 需要做一些基本假定, 把所研究的问题归结为一个统计问题, 然后用方差分析的方法进行解决.

在例 **??** 中, 我们只考察了一个因子, 称其为单因子试验. 通常, 在单因子试验中, 记因子为 A, 设其有 r 个水平, 记为 $A_1, A_2, \ldots, A_r$, 在每一水平下考察的指标可以看成一个总体, 现有 r 个水平, 故有 r 个总体, 假定:

1. 每一总体均为正态分布, 记为 $N(\mu_i, \sigma_i^2)$, $i = 1,\ldots,r$;
2. 各总体的方差相同, 记为 $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_r^2 = \sigma^2$;
3. 从每一总体中抽取的样本是相互独立的, 即所有的试验结果 y_{ij} 都相互独立.

这三个假定都可以用统计方法进行验证. 譬如, 利用正态性检验(7.4.3 节)验证 **??** 成立; 利用后面 **??** 的方差齐次性检验验证 **??** 成立; 而试验结果 y_{ij} 的独立性可由随机化实现, 这里的随机化是指所有试验按随机次序进行.

我们要做的工作是比较各水平下的均值是否相同, 即要对如下的一个假设进行检验,

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_r, \tag{8.1.1}$$

其备择假设为

$$H_1: \mu_1, \mu_2, \ldots, \mu_r \text{ 不全相等},$$

在不会引起误解的情况下, H_1 通常可省略不写.

如果 H_0 成立, 因子 A 的 r 个水平均值相同, 称因子 A 的 r 个水平间没有显著差异, 简称因子 A **不显著**; 反之, 当 H_0 不成立时, 因子 A 的 r 个水平均值不全相同, 这时称因子 A 的不同水平间有显著差异, 简称因子 A **显著**.

为对假设 (??) 进行检验, 需要从每一水平下的总体抽取样本, 设从第 i 个水平下的总体获得 m 个试验结果（简单起见, 这里先假设个水平下试验的重复数相同, 后面会看到, 重复数不同时的处理方式与此基本一致, 略有差异）, 记 y_{ij} 表示第 i 个总体的第 j 次重复试验结果. 共得到如下 $r\times m$ 个试验结果:

$$y_{ij},\ i=1,2,\ldots,r,\ j=1,2,\ldots,m,$$

其中 r 为水平数, m 为重复数, i 为水平编号, j 为重复编号.

在水平 A_i 下的试验结果 y_{ij} 与该水平下的指标均值 μ_i 一般总是有差距的, 记 $\varepsilon_{ij}=y_{ij}-\mu_i$, ε_{ij} 称为随机误差. 于是有

$$y_{ij}=\mu_i+\varepsilon_{ij} \tag{8.1.2}$$

(??) 式称为试验结果 y_{ij} 的**数据结构式**. 把三个假定用子数据结构式就可以写出单因子方差分析的统计模型：

$$\begin{cases} y_{ij}=\mu_i+\varepsilon_{ij},\ i=1,2,\ldots,r,\ j=1,2,\ldots,m; \\ \text{诸 } \varepsilon_{ij} \text{ 相互独立, 且都服从 } N(0,\sigma^2)\,. \end{cases} \tag{8.1.3}$$

为了能更好地描述数据, 常在方差分析中引入总均值与效应的概念. 称诸 μ_i 的平均（所有试验结果的均值的平均）

$$\mu=\frac{1}{r}(\mu_1+\cdots+\mu_r)=\frac{1}{r}\sum_{i=1}^{r}\mu_i \tag{8.1.4}$$

为总均值. 称第 i 水平下的均值 μ_i 与总均值 μ 的差

$$a_i=\mu_i-\mu,\quad i=1,2,\ldots,r \tag{8.1.5}$$

为因子 A 的第 i 水平的**主效应**, 简称为 A_i 的效应.

容易看出

$$\sum_{i=1}^{r}a_i=0, \tag{8.1.6}$$

$$\mu_i=\mu+a_i, \tag{8.1.7}$$

这表明第 i 个总体均值是由总均值与该水平的效应叠加而成的, 从而模型 (??) 可以改写为

$$\begin{cases} y_{ij}=\mu+a_i+\varepsilon_{ij},\quad i=1,2,\ldots,r,\ j=1,2,\ldots,m; \\ \sum\limits_{i=1}^{r}a_i=0; \\ \varepsilon_{ij} \text{ 相互独立, 且都服从 } N(0,\sigma^2)\,. \end{cases} \tag{8.1.8}$$

假设 (??) 可改写为

$$H_0:a_1=a_2=\cdots=a_r, \tag{8.1.9}$$

其备择假设为

$$H_1:a_1,a_2,\ldots,a_r \text{ 不全为 } 0.$$

8.1.2 平方和分解

一、试验数据

通常在单因子方差分析中可将试验数据列成如下表格形式.

表 8.1.2: 单因子方差分析试验数据

因子水平	试验数据				和	平均
A_1	y_{11}	y_{12}	$\cdots$	y_{1m}	T_1	$\bar{y}_1$
A_2	y_{21}	y_{22}	$\cdots$	y_{2m}	T_2	$\bar{y}_2$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$
A_r	y_{r1}	y_{r2}	$\cdots$	y_{rm}	T_r	$\bar{y}_r$
					T	$\bar{y}$

?? 中的最后二列的和与平均的含义如下：

$$T_i = \sum_{j=1}^{m} y_{ij},\ \bar{y}_i = \frac{T_i}{m} \quad i = 1,2,\ldots,r,$$

$$T_i = \sum_{i=1}^{r} T_i,\ \bar{y} = \frac{T}{r\cdot m} = \frac{T}{n},$$

$$n = r\cdot m = \text{总试验次数}.$$

二、组内偏差与组间偏差数据间是有差异的. 数据 y_{ij} 与总平均 $\bar{y}$ 间的偏差可用 $y_{ij}-\bar{y}$ 表示, 它可分解为两个偏差之和

$$y_{ij}-\bar{y} = (y_{ij}-\bar{y}_{i.}) + (\bar{y}_{i.}-\bar{y}) \tag{8.1.10}$$

记

$$\bar{\varepsilon}_{i.} = \frac{1}{m}\sum_{j=1}^{m}\varepsilon_{ij}, \quad \bar{\varepsilon} = \frac{1}{r}\sum_{i=1}^{r}\bar{\varepsilon}_i = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{m}\varepsilon_{ij}.$$

由于

$$y_{ij}-\bar{y}_{i.} = (\mu_i+\varepsilon_{ij}) - (\mu_i+\bar{\varepsilon}_i) = \varepsilon_{ij}-\bar{\varepsilon}_i, \tag{8.1.11}$$

所以 $y_{ij}-\bar{y}_{i.}$ 仅反映组内数据与组内平均的随机误差, 称为**组内偏差**; 而

$$\bar{y}_{i.}-\bar{y} = (\mu_i+\bar{\varepsilon}_{i.}) - (\mu+\bar{\varepsilon}_i) = a_i+\bar{\varepsilon}_{i.}-\bar{\varepsilon}, \tag{8.1.12}$$

$\bar{y}_{i.}-\bar{y}$ 除了反映随机误差外，还反映了第 i 个水平的效应，称为组间偏差，

三、偏差平方和及其自由度

在统计学中，把 k 个数据 $y_1,\ldots,y_k$ 分别对其均值 $\bar{y} = (y_1+\cdots+y_k)/k$ 的偏差平方和

$$Q = (y_1-\bar{y})^2+\cdots+(y_k-\bar{y})^2 = \sum_{i=1}^{k}(y_i-\bar{y})^2$$

称为 k 个数据的**偏差平方和**，有时简称**平方和**. 偏差平方和常用来度量若干个数据集中或分散的程度，它是用来度量若干个数据间差异（即波动）的大小的一个重要的统计量.

在构成偏差平方和 Q 的 k 个偏差 $y_1-\bar{y},\ldots,y_k-\bar{y}$ 间有一个恒等式

$$\sum_{i=1}^{k}(y_i-\bar{y}) = 0$$

这说明在 Q 中独立的偏差只有 $k-1$ 个. 在统计学中把平方和中独立偏差个数称为该平方和的**自由度**，常记为 f ，如 Q 的自由度为 $f_Q = k-1$. 自由度是偏差平方和的一个重要参数.

四、总平方和分解公式

各 y_{ij} 间总的差异大小可用**总偏差平方和** S_T 表示，

$$S_T = \sum_{i=1}^{r}\sum_{j=1}^{m}(y_{ij}-\bar{y})^2, \quad f_T = n-1, \tag{8.1.13}$$

仅由随机误差引起的数据间的差异可以用组内偏差平方和表示，也称为误差偏差平方和，记为 S_e

$$S_e = \sum_{i=1}^{r}\sum_{j=1}^{m}(y_{ij}-\bar{y}_{i.})^2, \quad f_e = r(m-1) = n-r. \tag{8.1.14}$$

由于组间差异除了随机误差外，还反映了效应间的差异，故由效应不同引起的数据差异可用**组间偏差平方和**表示，也称为因子 A 的**偏差平方和**，记为 S_A：

$$S_A = m\sum_{i=1}^{r}(\bar{y}_{i.}-\bar{y})^2, \quad f_A = r-1 \tag{8.1.15}$$

定理 8.1.1. *在上述符号下，总平方和 S_T 可以分解为因子平方和 S_A 与误差平方和 S_e。之和，其自由度也有相应分解公式，具体为：*

$$S_T = S_A + S_e, \quad f_T = f_A + f_e \tag{8.1.16}$$

(??) 式通常称为总平方和分解式.

证明： 注意到

$$\sum_{i=1}^{r}\sum_{j=1}^{m}(y_{ij}-\bar{y}_{i.})(\bar{y}_{i.}-\bar{y}) = \sum_{i=1}^{r}\left[(\bar{y}_{i.}-\bar{y})\sum_{i=1}^{m}(y_{ij}-\bar{y}_{i.})\right] = 0 \tag{8.1.17}$$

故有

$$\begin{aligned} S_T &= \sum_{i=1}^{r}\sum_{j=1}^{m}(y_{ij}-\bar{y})^2 = \sum_{i=1}^{r}\sum_{i=1}^{m}[(y_{ij}-\bar{y}_{i.})+(\bar{y}_{i.}-\bar{y})]^2 \\ &= S_e + S_A + 2\sum_{i=1}^{r}\sum_{j=1}^{m}(y_{ij}-\bar{y}_{i.})+(\bar{y}_{i.}-\bar{y}) = S_e + S_A, \end{aligned}$$

诸自由度间的等式是显然的.

8.1.3 检验方法

偏差平方和 Q 的大小与数据个数（或自由度）有关，一般说来，数据越多，其偏差平方和越大. 为了便于在偏差平方和间进行比较，统计上引入了均方和的概念，它定义为

$$MS = Q/f_Q,$$

其意为平均每个自由度上有多少平方和，它比较好地度量了一组数据的离散程度.

如今要对因子平方和 SA 与误差平方和 S2 之间进行比较，用其均方和

$$MS_A = S_A/f_A, \quad MS_e = S_e/f_e$$

进行比较更为合理，因为均方和排除了自由度不同所产生的干扰. 故用

$$F = \frac{MS_A}{MS_e} = \frac{S_A/f_A}{S_e/f_e} \tag{8.1.18}$$

作为检验 H_0 的统计量，为给出检验拒绝域，我们需要如下定理：

定理 8.1.2. *在单因子方差分析模型 (??) 及前述符号下，有*

1. $S_\varepsilon/\sigma^2 \sim \chi^2(n-r)$, *从而* $E(S_e) = (n-r)\sigma^2$

2. $E(S_A)=(r-1)\sigma^2+m\sum_{i=1}^{r}a_i^2$, 进一步, 若 H_0 成立, 则有 $S_A/\sigma^2\sim\chi^2(r-1)$;
3. S_A 与 S_e 独立 .

证明： 由于 (??) 和 (??), $S_e=\sum_{i=1}^{r}\sum_{j=1}^{m}(\varepsilon_{ij}-\bar{\varepsilon_{i.}})^2$, 在单因子方差分析模型 (??) 下, 我们知道, 诸 ε_{ij}, $i=1,2,\ldots,r$, $j=1,2,\ldots,m$ 独立同分布于 $N(0,\sigma^2)$, 由定理 ?? 知, $\frac{1}{\sigma^2}\sum_{j=1}^{m}(\varepsilon_{ij}-\bar{\varepsilon}_{i.})^2$, $i=1,2,\ldots,r$, 相互独立, 其共同分布为 $\chi^2(m-1)$, 由卡方分布的可加性, 有 $\frac{S_e}{\sigma^2}\sim\chi^2(n-r)$, 这给出 $E(S_e/\sigma^2)=n-r=f_e$, ??得证.

类似地, 由 (??) 和 (??), 有

$$S_A=m\sum_{i=1}^{r}(a_i+\varepsilon_{i.}-\bar{\varepsilon})^2.$$

由定理 ?? 知, 对每个 i, 平方和 $\sum_{j=1}^{m}(\varepsilon_{ij}-\bar{\varepsilon}_{i.})^2$ 与均值 $\bar{\varepsilon}_i$ 独立, 从而 $\bar{\varepsilon}_{1.},\bar{\varepsilon}_{2.},\ldots,\bar{\varepsilon}_{r.}$ 与 S_e 独立, 而 S_A 只是, $\bar{\varepsilon}_{1.},\bar{\varepsilon}_{2.},\ldots,\bar{\varepsilon}_{r.}$ 的函数, 由此 ?? 得证.

在模型 ?? 下, S_A 的期望是

$$E(S_A)=m\sum_{i=1}^{r}a^2+E\Big[m\sum_{i=1}^{r}(\bar{\varepsilon}_{i.}-\bar{\varepsilon})^2\Big],$$

由于诸误差均值 $\bar{\varepsilon}_{1.},\bar{\varepsilon}_{2.},\ldots,\bar{\varepsilon}_{r.}$ 独立同分布于 $N(0,\sigma^2/m)$, 从而由诸误差均值组成的偏差平方和除以 σ^2/m 服从卡方分布, 即

$$\frac{1}{\sigma^2}\sum_{i=1}^{r}m\,(\bar{\varepsilon}_{i},-\bar{\varepsilon})^2-\chi^2(r-1).$$

于是, $E\big[\sum_{i=1}^{r}m(\bar{\varepsilon}_{i.}-\bar{\varepsilon})\big]$ 在 H_0 成立下, $S_A/\sigma^2\sim\chi^2(r-1)$, 这就完成了 ?? 的证明.

8.2 多重比较

8.3 方差齐次检验

8.4 一元线性回归

8.5 一元非线性回归