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

R E-Book

普通高等教育“十五”国家级规划教材

目　录

第 1 章 方差分析与回归分析

1.1 方差分析

1.1.1 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 1.1.1: 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 1.1.1: 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 1.1.1 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 1. 成立; 利用后面 1.3 的方差齐次性检验验证 2. 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,

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_r, \tag{1.1.1}$$

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 (1.1.1) 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

$$y_{ij}=\mu_i+\varepsilon_{ij} \tag{1.1.2}$$

(1.1.2) 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：

$$\begin{cases} y_{ij}=\mu_i+\varepsilon_{ij},\ i=1,2,\ldots,r,\ j=1,2,\ldots,m; \\ \text{诸 } \varepsilon_{ij} \text{ 相互独立, 且都服从 } N(0,\sigma^2)\ . \end{cases} \tag{1.1.3}$$

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)

$$\mu=\frac{1}{r}(\mu_1+\cdots+\mu_r)=\frac{1}{r}\sum_{i=1}^{r}\mu_i \tag{1.1.4}$$

为总均值. 称第 i 水平下的均值 μ_i 与总均值 μ 的差

$$a_i=\mu_i-\mu,\quad i=1,2,\ldots,r \tag{1.1.5}$$

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

$$\sum_{i=1}^{r}a_i=0, \tag{1.1.6}$$

$$\mu_i=\mu+a_i, \tag{1.1.7}$$

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 (1.1.3) 可以改写为

$$\begin{cases} y_{ij}=\mu+a_i+\varepsilon_{ij}, \quad i=1,2,\ldots,r,\ j=1,2,\ldots,m; \\ \sum\limits_{i=1}^{r}a_i=0; \\ \varepsilon_{ij} \text{ 相互独立, 且都服从 } N(0,\sigma^2)\ . \end{cases} \tag{1.1.8}$$

假设 (1.1.1) 可改写为

$$H_0: a_1=a_2=\cdots=a_r, \tag{1.1.9}$$

其备择假设为

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

1.1.2 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 1.1.2: 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	$\bar{y}$

1.1.2 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

$$y_{ij} - \bar{y} = (y_{ij} - \bar{y}_{i.}) + (\bar{y}_{i.} - \bar{y}) \tag{1.1.10}$$

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

$$y_{ij} - \bar{y}_{i.} = (\mu_i + \varepsilon_{ij}) - (\mu_i + \bar{\varepsilon}_i) = \varepsilon_{ij} - \bar{\varepsilon}_i, \tag{1.1.11}$$

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

$$\bar{y}_{i.} - \bar{y} = (\mu_i + \bar{\varepsilon}_{i.}) - (\mu + \bar{\varepsilon}_i) = a_i + \bar{\varepsilon}_{i.} - \bar{\varepsilon}, \tag{1.1.12}$$

$\bar{y}_{i.} - \bar{y}$ 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，

$$S_T = \sum_{i=1}^{r}\sum_{j=1}^{m}(y_{ij}-\bar{y})^2, \quad f_T = n-1, \tag{1.1.13}$$

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

$$S_e = \sum_{i=1}^{r}\sum_{j=1}^{m}(y_{ij}-\bar{y}_{i.})^2, \quad f_e = r(m-1) = n-r. \tag{1.1.14}$$

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；

$$S_A = m\sum_{i=1}^{r}(\bar{y}_{i.}-\bar{y})^2, \quad f_A = r-1 \tag{1.1.15}$$

定理 1.1.1

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：

$$S_T = S_A + S_e, \quad f_T = f_A + f_e \tag{1.1.16}$$

(1.1.16) 式通常称为总平方和分解式. ♡

证明： 注意到

$$\sum_{i=1}^{r}\sum_{j=1}^{m}(y_{ij}-\bar{y}_{i.})(\bar{y}_{i.}-\bar{y}) = \sum_{i=1}^{r}\left[(\bar{y}_{i.}-\bar{y})\sum_{i=1}^{m}(y_{ij}-\bar{y}_{i.})\right] = 0 \tag{1.1.17}$$

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.

1.1.3 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

$$F = \frac{MS_A}{MS_e} = \frac{S_A/f_A}{S_e/f_e} \tag{1.1.18}$$

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 1.1.2

在单因子方差分析模型 (1.1.8) 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 (1.1.11) 和 (1.1.14), $S_e = \sum_{i=1}^{r}\sum_{j=1}^{m}(\varepsilon_{ij} - \bar{\varepsilon_{i.}})^2$, 在单因子方差分析模型 (1.1.8) 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 $E(S_e/\sigma^2) = n - r = f_e$, 1.得证.

类似地，由 (1.1.12) 和 (1.1.15)，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 3. 得证.

在模型 1.1.8 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 2. 的证明.

1.2 多重比较

1.3 方差齐次检验

1.4 一元线性回归

1.5 一元非线性回归