常用统计方法

王玲玲　周纪芗

华东师范大学出版社

常 用 统 计 方 法

王玲玲　周纪芗

华东师范大学出版社
1994年·上海

(沪)新登字第201号

常用统计方法

王玲玲　周纪芗

华东师范大学出版社出版发行
(上海中山北路3663号)
邮政编码：200062
新华书店上海发行所经销　　江苏句容排印厂印刷
开本：850×1168　1/32　印张：11.5　字数：290千字
1994年9月第一版　　1994年9月第一次印刷
印数：001—3,000本

ISBN 7-5617-1196-4/O·039　　定　价：11.30元

前　言

随着生产力的发展、科技的进步，许多工程技术人员、科研人员及经济管理人员越来越重视各种数理统计方法在各自领域中的应用．本书旨在深入浅出地介绍一些在工农业生产、科学实验和研究、经济活动中有广泛应用价值的数理统计方法，其中包括方差分析、正交试验设计、回归分析、回归设计、协方差分析及各种常用的多元分析方法．这些方法在质量管理、经济预测、农业、水利、生物、医学、地质、气象、地震预报、自动控制、教育、体育研究等方面都有广泛的应用．

对每一种统计方法，我们力求交代清楚实际背景、统计思想、数学模型、解决问题的思路与具体操作步骤，并用例子加以说明．本书起点较低，稍有概率统计基础知识的读者都不难掌握这些方法．

作为未来建设主力军的大学生学习掌握这些方法，将会在今后工作中发挥更大作用．本书可作为理、工、农、医、经济类高等学校的教材，也可作为工程技术人员的培训教材．每一章后附有一定数量的习题，以便读者通过练习加深对内容的理解．

本书的主要内容在华东师范大学数学、地理、电子、计算机、化学、经济、教育等专业大专生、本科生、研究生的课程中多次讲授．所用的教材经过几次修改，得到不断完善．在本书定稿时，我们对所给的实例，重新进行了验算．由于本书所介绍的各种统计方法有相对的独立性，所以可以根据不同的教学对象和要求，灵活选择其中的某些章节讲授，灵活掌握深浅程度．对数学要求较低的专业可略去数学模型及公式的推导，重点讲授方法的实际背景及解题的具体操作步骤．

由于编者水平有限，书中难免有不妥之处，恳请广大读者批评指正。

王玲玲　周纪芗

1993年于华东师大

目　录

第一章 方差分析

§1.1 引 言

在工农业生产和科学研究中，经常要分析各种因素对研究对象某些特性值的影响. 例如，在工业生产中往往要考察几种不同原料对产品质量有否明显影响；几位检验员检查同一型号的产品，要了解他们的检验技术有无明显不同；在农业生产中我们要考察不同品种、施肥种类、施肥量等对某种作物亩产量的影响；在化学实验中要分析反应温度、反应时间、原料成分、原料用量等对某种物质得率的影响……. 为了分析这些因素对特性值的影响，就必须让这些因素改变各种不同状态进行试验或考察，并对所得结果——数据进行科学的分析. 方差分析就是采用数理统计方法对所得结果进行分析，以鉴别各种因素对研究对象的某些特性值影响大小的一种有效方法.

为方便起见，今后我们把研究对象的特性值，即试验(其涵义包括调查、收集等)结果，如产量，某些质量指标称为试验指标，简称指标，常用 y 表示. 在试验中要加以考察而改变状态的因素称为因子，常用 A，B，C 等大写英文字母表示. 因子在试验中所取的各种不同状态称为因子的水平，常用 A_1，A_2，…，A_r 等表示，其中 r 称为因子 A 的水平数.

例 1.1 为寻求适应某地区的高产油菜品种，今选了五种不同品种进行试验，每一品种在四块条件相同的试验田上试种，其它施肥等田间管理措施完全一样，表 1.1 为每一品种下每一块田的亩产量和每一品种下四块田的平均亩产量(每一品种下，四个原始数据平均值).

表 1.1 数据表

品种 / 田块	A_1	A_2	A_3	A_4	A_5
1	256	244	250	288	206
3	222	300	277	280	212
3	280	290	230	315	220
4	298	275	322	279	272
平均亩产	264	277.25	269.75	285.50	212.[illegible]

要根据这些数据分析不同油菜品种对平均亩产影响是否显著.

从平均亩产来看，好像不同品种对亩产有一定影响. 但仔细分析一下数据，问题就不那么简单. 可以看到，在同一种品种下，四块不同田块的亩产也不完全一样，试验时已考虑到田块及其它条件一样，产生这种差异的原因是由于试验过程中各种偶然性因素的干扰及称量误差等所致，这一类误差称为试验误差，由于试验误差的存在使平均亩产中也含有试验误差. 因此对于不同品种下平均亩产的差异应作仔细的分析，这差异单纯是由误差引起的，还是由于油菜品种不同而引起的. 如果平均亩产的差异单纯是由误差引起的，那么我们认为油菜的五种不同品种对亩产没有显著影响. 若用因子 A 表示油菜品种(它共有五个水平)，则可简称因子 A 不显著. 如果不同水平下平均亩产的不同，除了误差影响外，主要是由于水平不同所造成的，那么我们就认为因子 A 的不同水平对亩产有显著影响，简称因子 A 显著. 方差分析就是通过对试验结果的分析去判断因子是否显著的一种统计方法.

例 1.1 只考察一个因子对指标的影响，这种试验称为单因子试验，相应的方差分析就称为单因子方差分析，若一个试验中同时考察两个因子，则相应的试验称为两因子试验，这时所作的方差分析称为两因子方差分析，在多因子试验中要考察的因子多于两个，

相应的方差分析称为多因子方差分析.

单因子方差分析是最简单的,有时结论也是一目了然的,但正因为简单,所以对理解方差分析的思想和方法有帮助,因而我们先介绍单因子方差分析.

§1.2 单因子方差分析

一、模型

考虑一个因子 A 取 r 个水平,分析这 r 个不同水平对指标 y 的影响. 为此在每个水平 A_i 下重复做 m 次试验, $i=1, 2, \cdots, r$, 共得 $n=r\times m$ 个数据, 见数据表 1.2. 表中 y_{ij} 表示在 A_i 水平下第 j 次试验结果.

表 1.2 原始数据表

重复 \ 水平	A_1	A_2	$\cdots$	A_r
1	y_{11}	y_{21}	$\cdots$	y_{r1}
2	y_{12}	y_{22}	$\cdots$	y_{r2}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
m	y_{1m}	y_{2m}	$\cdots$	y_{rm}

一般我们假定在 A_i 水平下指标 $y_{ij}\sim N(\mu_i, \sigma^2)$, $j=1, 2, \cdots, m$, $i=1, 2, \cdots, r$. 这表明各水平下指标 y_{ij} 是服从正态分布的随机变量. 在同一水平下,它们的平均值为 μ_i, 它们可能相同,也可能不同. 若一切 μ_i 都相同,这就表示各水平下指标值无显著差异, 否则就有显著差异. 这里要求各水平下指标值的波动大小是一致的,即在不同水平 A_i 下,各次试验中所得随机变量 y_{ij} 的方差 σ^2 是相等的. 由此可见,在单因子方差分析中就是要通过对数据 y_{ij} 的分析去判断 $\mu_1, \mu_2, \cdots, \mu_r$ 是否全部相同, 即要检验假设

$$H_0: \mu_1=\mu_2=\cdots=\mu_r$$

是否成立.

分析 A_i 水平下数据 y_{ij}，它取值的平均水平为 μ_i，而所以在 m 次重复试验中有各不相同的取值，是由于试验过程中随机误差引起的. 如果我们用一个随机变量 ε_{ij} 表示第 i 个水平下第 j 次试验的随机误差，那末数据 y_{ij} 就有如下结构形式：

$$y_{ij}=\mu_i+\varepsilon_{ij},\quad i=1,2,\cdots,r,\ j=1,2,\cdots,m. \tag{1.1}$$

由于 $y_{ij}\sim N(\mu_i,\sigma^2)$，所以 $\varepsilon_{ij}\sim N(0,\sigma^2)$. 为了今后讨论方便，我们引入如下记号. 令

$$\mu=\frac{1}{r}\sum_{i=1}^{r}\mu_i, \tag{1.2}$$

$$a_i=\mu_i-\mu,\ i=1,2,\cdots,r, \tag{1.3}$$

称 μ 为一般平均，a_i 为因子 A 的第 i 个水平的效应. 它的大小反映了该水平相对于一般平均的差别大小，显然

$$\sum_{i=1}^{r}a_i=\sum_{i=1}^{r}(\mu_i-\mu)=\sum_{i=1}^{r}\mu_i-r\mu=0. \tag{1.4}$$

利用 μ 和 a_i，数据的结构形式可以表示成：

$$y_{ij}=\mu+a_i+\varepsilon_{ij},\ i=1,2,\cdots,r,\ j=1,2,\cdots,m. \tag{1.5}$$

将公式(1.4)，(1.5)与关于 ε_{ij} 的分布合并写在一起，就可以得到单因子方差分析中数学模型：

$$\begin{cases} y_{ij}=\mu+a_i+\varepsilon_{ij},\quad i=1,2,\cdots,r,\ j=1,2,\cdots,m,\\ \sum\limits_{i=1}^{r}a_i=0,\\ \varepsilon_{ij}\sim N(0,\sigma^2),\text{且相互独立},\quad i=1,2,\cdots,r,\\ \qquad\qquad j=1,2,\cdots,m. \end{cases} \tag{1.6}$$

并用这些数据检验假设

$$H_0:\ a_1=a_2=\cdots=a_r=0. \tag{1.7}$$

根据公式(1.3)，可知此假设与 $\mu_1=\mu_2=\cdots=\mu_r$ 的假设是一致的. 关于对 ε_{ij} 相互独立的要求，即要求各次试验是独立进行的，这样就能保证每次试验的随机误差 ε_{ij} 是相互独立的随机变量，$i=1,2,\cdots,r,\ j=1,2,\cdots,m$.

二、检验统计量，偏差平方和的分解

为寻求检验假设 H_0 的统计量，我们可从分析数据 y_{ij} 的差异原因着手. 各 y_{ij} 取值不同，如前分析，主要原因有两个，一是可能 A 取不同水平所引起，二是有随机误差. 我们希望能用具体的量来刻划这些差异，为此引入一些符号，令

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

为 A_i 水平下数据的和及数据的平均值. 并设

$$T=\sum_{i=1}^{r}\sum_{j=1}^{m} y_{ij}=\sum_{i=1}^{r} T_i,\qquad \bar{y}=\frac{T}{rm}=\frac{T}{n}$$

为所有数据的和及所有数据的平均值.

数据总的差异可以用总偏差平方和 S_T 这个量来表示:

$$S_T=\sum_{i=1}^{r}\sum_{j=1}^{m}(y_{ij}-\bar{y})^2 \tag{1.8}$$

总的偏差来自两个方面，为此我们可以对总偏差平方和 S_T 进行分解.

$$S_T=\sum_{i=1}^{r}\sum_{j=1}^{m}(y_{ij}-\bar{y})^2=\sum_{i=1}^{r}\sum_{j=1}^{m}(y_{ij}-\bar{y}_i+\bar{y}_i-\bar{y})^2$$

$$=\sum_{i=1}^{r}\sum_{j=1}^{m}(y_{ij}-\bar{y}_i)^2+\sum_{i=1}^{r}\sum_{j=1}^{m}(\bar{y}_i-\bar{y})^2,$$

其中交叉项为零. 若记

$$S_e=\sum_{i=1}^{r}\sum_{j=1}^{m}(y_{ij}-\bar{y}_i)^2, \tag{1.9}$$

$$S_A=\sum_{i=1}^{r}\sum_{j=1}^{m}(\bar{y}_i-\bar{y})^2=\sum_{i=1}^{r} m(\bar{y}_i-\bar{y})^2. \tag{1.10}$$

可以看到，前者反映同一水平下数据 y_{ij} 与其平均值 $\bar{y}_i$ 的差异，它是试验误差引起的；后者是不同水平下数据的平均值与所有数据的总平均值之间的偏差平方和，它包含了因子 A 取不同水平引起的数据差异，也包含了试验误差对它的影响. 为清楚地说明此问题，可以利用数据结构式(1.6)进行分析. 由(1.6)式可得:

$$\bar{y}_i=\frac{1}{m}\sum_{j=1}^{m}(\mu+a_i+\varepsilon_{ij})=\mu+a_i+\bar{\varepsilon}_i,$$

$$\bar{y}=\frac{1}{rm}\sum_{i=1}^{r}\sum_{j=1}^{m}(\mu+a_i+\varepsilon_{ij})=\mu+\bar{\varepsilon},$$

其中 $\bar{\varepsilon}_i=\frac{1}{m}\sum_{j=1}^{m}\varepsilon_{ij}$，$\bar{\varepsilon}=\frac{1}{rm}\sum_{i=1}^{r}\sum_{j=1}^{m}\varepsilon_{ij}$，因而有

$$S_e=\sum_{i=1}^{r}\sum_{j=1}^{m}(y_{ij}-\bar{y}_i)^2=\sum_{i=1}^{r}\sum_{j=1}^{m}[(\mu+a_i+\varepsilon_{ij})-(\mu+a_i+\bar{\varepsilon}_i)]^2$$

$$=\sum_{i=1}^{r}\sum_{j=1}^{m}(\varepsilon_{ij}-\bar{\varepsilon}_i)^2,$$

$$S_A=\sum_{i=1}^{r}m(\bar{y}_i-\bar{y})^2=\sum_{i=1}^{r}m[(\mu+a_i+\bar{\varepsilon}_i)-(\mu+\bar{\varepsilon})]^2$$

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

可见 S_e 中只单纯含有试验误差的影响，由于 $\varepsilon_{ij}\sim N(0,\ \sigma^2)$，且相互独立，所以有 $\frac{S_e}{\sigma^2}\sim\chi^2(r(m-1))$. S_A 中既有因子 A 效应 a_i 的影响，也有试验误差的影响，由于 $\bar{\varepsilon}_i\sim N\left(0,\ \frac{\sigma^2}{m}\right)$，所以在假设 H_0 为真时，$\frac{S_A}{\sigma^2}\sim\chi^2(r-1)$.

通过对总偏差平方和 S_T 的分解得

$$S_T=S_e+S_A, \tag{1.11}$$

此式称为平方和分解公式，不难找到检验假设 H_0 的统计量

$$F=\frac{\dfrac{S_A}{f_A}}{\dfrac{S_e}{f_e}}, \tag{1.12}$$

其中 $f_A=r-1$，是假设(1.7)成立时 S_A 的自由度，$f_e=r(m-1)$ 是 S_e 的自由度. 当 F 取值较大时，说明在 S_A 中，因子 A 的效应 a_i 的影响不可忽略，因而认为假设 H_0 不成立，即因子 A 取 r 个不同水平对指标 y 有显著影响，简称因子 A 显著　若 F 取值较小，则 S_A 中效应 a_i 的影响不大，它主要是试验误差引起，因而可以认为因子 A 不显著. 为求检验的临界值，必须知道在假设 H_0 为真时统计量 F 的分布，而此关键是要考虑 S_A 和 S_e 的独立性，为此引入一个分解定理. 此定理又称 χ^2 分解定理或柯赫伦

(Cochran)定理

分解定理（Cochran 定理） 设 $x_1, x_2, \cdots, x_n$ 为 n 个相互独立的 $N(0, 1)$ 变量，$Q=\sum_{i=1}^{n} x_i^2$ 为 $\chi^2(n)$ 变量，若 $Q=Q_1+Q_2+\cdots+Q_k$，其中 Q_i 为某些正态变量的平方和，这些正态变量分别是 $x_1, x_2, \cdots, x_n$ 的线性组合，其自由度为 f_i，则诸 Q_i 相互独立，且服从 $\chi^2(f_i)$ 分布的充要条件是：

$$f_1+f_2+\cdots+f_k=n.$$

证 必要性. 若 $Q_1, Q_2, \cdots, Q_k$ 相互独立，且 $Q_i\sim\chi^2(f_i)$，$i=1, 2, \cdots, k$，则由 χ^2-分布的可加性知

$$Q=\sum_{i=1}^{k} Q_i\sim\chi^2\left(\sum_{i=1}^{k} f_i\right),$$

又 $Q\sim\chi^2(n)$，所以

$$n=f_1+f_2+\cdots+f_k.$$

充分性. 设 z_{ij} 为正态变量，$i=1, 2, \cdots, k$，$j=1, 2, \cdots, m_i$，且 $Q_i=\sum_{j=1}^{m_i} z_{ij}^2$，由假定在 $z_{i1}, z_{i2}, \cdots, z_{im_i}$ 中必可选出 f_i 个 z_{ij}，而其余的可由 f_i 个线性表示，不妨设 $z_{i,f_i+1}\cdots, z_{im_i}$ 可由 $z_{i1}, \cdots, z_{if_i}$ 线性表示，将这些关系式代入 Q_i 后即得 Q_i 为 $z_{i1}, \cdots, z_{if_i}$ 的一个非负二次型，由二次型的理论可知，将此二次型标准化后得

$$Q_i=\sum_{j=1}^{f_i} b_{ij}\tilde{z}_{ij}^2,$$

其中 $\tilde{z}_{ij}$ 是 $z_{i1}, \cdots, z_{if_i}$ 的线性组合. 又由于 z_{ij} 是 $x_1, x_2, \cdots, x_n$ 的线性组合，故 $\tilde{z}_{ij}$ 为独立正态变量 $x_1, x_2, \cdots, x_n$ 的线性组合，所以它们仍为正态变量；$b_{ij}=+1$ 或 -1，从而

$$Q=\sum_{i=1}^{k}\sum_{j=1}^{f_i} b_{ij}\tilde{z}_{ij}^2=\sum_{i=1}^{n} x_i^2.$$

由于 Q 是正定的；又 $\sum_{i=1}^{k} f_i=n$，故 $\tilde{z}_{ij}$ 共有 n 个，且一切 b_{ij} 全取为 $+1$. 将 $\tilde{z}_{ij}$ 重新编号成 $\tilde{z}_1, \tilde{z}_2, \cdots, \tilde{z}_n$，则

$$Q=\sum_{i=1}^{n} x_i^2=\sum_{i=1}^{n}\tilde{z}_i^2.$$

从而可知由 $x_1, x_2, \cdots, x_n$ 到 $\tilde{z}_1, \tilde{z}_2, \cdots, \tilde{z}_n$ 的线性变换是正交变换，$\tilde{z}_1, \tilde{z}_2, \cdots, \tilde{z}_n$ 仍是正态变量，且由正交变换性质知有

$$E\tilde{z}_i=0,$$

$$\operatorname{cov}(\tilde{z}_i, \tilde{z}_j)=\begin{cases}0, & i\neq j,\\ 1, & i=j,\end{cases} \quad i, j=1, 2, \cdots, n.$$

这就说明 $\tilde{z}_1, \tilde{z}_2, \cdots, \tilde{z}_n$ 也是相互独立的 $N(0, 1)$ 变量，故诸 Q_i 为相互独立的 $\chi^2(f_i)$ 变量，定理证毕.

在总偏差平方和 S_T 的分解式 $S_T=S_e+S_A$ 中，由于 $y_{ij}\sim N(\mu_i, \sigma^2)$，所以在假设 H_0 成立下有 $\frac{S_T}{\sigma^2}=\sum_{i=1}^{r}\sum_{j=1}^{m}(y_{ij}-\bar{y})^2/\sigma^2\sim\chi^2(n-1)$，$S_A^2/\sigma^2=\sum_{i=1}^{r}m(\bar{y}_i-\bar{y})^2/\sigma^2\sim\chi^2(r-1)$. 不管 H_0 是否成立，总有 $S_e^2/\sigma^2=\sum_{i=1}^{r}\sum_{j=1}^{m}(y_{ij}-\bar{y}_i)^2/\sigma^2\sim\chi^2(r(m-1))$. 又有 $f_e+f_A=r(m-1)+r-1=rm-1=n-1=f_T$，所以由分解定理可知 S_e 与 S_A 独立，从而 $F=\frac{S_A/f_A}{S_e/f_e}$ 在假设 H_0 成立下服从 $F(r-1, r(m-1))$ 分布，而检验假设 H_0 的拒绝域为

$$F=\frac{S_A/f_A}{S_e/f_e}\geqslant F_\alpha(r-1, r(m-1)),$$

其中 α 为显著性水平，$F_\alpha(r-1, r(m-1))$ 是自由度为 $r-1$，$r(m-1)$ 的 F 分布 α 上侧分位数，其数值可查附表3. α 越小，拒绝 H_0 的把握越大(因 α 是在 H_0 成立时拒绝 H_0 的概率)，因子 A 的显著性越高. 一般取 $\alpha=0.05$ 时，称因子 A 显著，记 *；$\alpha=0.01$ 时，称因子 A 高度显著，记 **；$\alpha=0.1$ 时，称因子 A 一般显著，记 ⊛.

三、计算公式的简化与表格化

综上所述要分析因子 A 取 r 个水平对指标 y 是否有显著影响，可以利用检验统计量 F 对假设 H_0 作检验，为计算 F 的观察值，必需计算 S_A, S_e，或 S_T 中任意两个，公式(1.8)、(1.9)、(1.10)已给出了它们的计算公式，但可以简化

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

$$=\sum_{i=1}^{r}\sum_{j=1}^{m}y_{ij}^2-\frac{T^2}{n},\tag{1.13}$$

$$S_A=\sum_{i=1}^{r}m(\bar{y}_i-\bar{y})^2=\sum_{i=1}^{r}m\bar{y}_i^2-n\bar{y}^2=\sum_{i=1}^{m}\frac{T_i^2}{m}-\frac{T^2}{n},\tag{1.14}$$

$$S_e=S_T-S_A.\tag{1.15}$$

具体计算过程可按下列步骤进行：

(1) 在原始数据表中下面加二行，分别计算 A_i 水平下数据 y_{ij} 的和 T_i 及其平方 T_i^2.

(2) 计算 $T=\sum_{i=1}^{r}T_i$, $\sum_{i=1}^{r}\sum_{j=1}^{m}y_{ij}^2$, $\sum_{i=1}^{r}T_i^2$, $CT=\frac{T^2}{n}$.

(3) 按公式(1.13), (1.14), (1.15)依次计算出 S_T, S_A, S_e.

(4) 列方差分析表(表1.3)，并给出结论.

表1.3 方差分析表

来 源	平方和 S	自由度 f	均方和 V	F 比	显著性
因子 A	$S_A=\sum_{i=1}^{r}T_i^2/m-CT$	$r-1$	$V_A=S_A/(r-1)$	$F=V_A/V_e$	
误差 e	$S_e=S_T-S_A$	$n-r$	$V_e=S_e/(n-r)$		
总 和	$S_T=\sum_{i=1}^{r}\sum_{j=1}^{m}y_{ij}^2-CT$	$n-1$			

显著性是根据 $F_\alpha(r-1,\ n-r)$ 值与 F 值比较后作出结论，$F_\alpha(r-1,\ n-r)$ 可查 F 分布的 α 上侧分位数表.

下面我们来完成对例1.1的分析，数据表见表1.4.

本例中，$r=5$, $m=4$, $n=20$, $T=\sum_{i=1}^{5}T_i=5236$,

$$\sum_{i=1}^{5}\sum_{j=1}^{4}y_{ij}^2=1395472,\ CT=\frac{T^2}{n}=\frac{(5236)^2}{20}=1370784.8,$$

$$S_T=\sum_i\sum_j y_{ij}^2-CT=1395472-1370784.8=24687.2,$$