

第十一章 方差分析

我们已经作过两个总体均值的假设检验，如两台机床生产的零件尺寸是否相等，病人和正常人的某个生理指标是否一样。如果把这类问题推广一下，要检验两个以上总体的均值彼此是否相等，仍然用以前介绍的方法是很难做到的。而你在实际生产和生活中可以举出许多这样的问题：从用几种不同工艺制成的灯泡中，各抽取了若干个测量其寿命，要推断这几种工艺制成的灯泡寿命是否有显著差异；用几种化肥和几个小麦品种在若干块试验田里种植小麦，要推断不同的化肥和品种对产量有无显著影响。

可以看到，为了使生产过程稳定，达到优质、高产，需要对影响产品质量的因素进行分析，找出有显著影响的那些因素，除了从机理方面进行研究外，常常要作许多试验，对结果作分析、比较，寻求规律。用数理统计分析试验结果、鉴别各因素对结果影响程度的方法称为方差分析 (Analysis Of Variance)，记作 ANOVA。

人们关心的试验结果称为**指标**，试验中需要考察、可以控制的条件称为**因素或因子**，因素所处的状态称为**水平**。上面提到的灯泡寿命问题是单因素试验，小麦产量问题是双因素试验。处理这些试验结果的统计方法就称为单因素方差分析和双因素方差分析。

§1 单因素方差分析

只考虑一个因素 A 对所关心的指标的影响， A 取几个水平，在每个水平上作若干个试验，试验过程中除 A 外其它影响指标的因素都保持不变（只有随机因素存在），我们的任务是从试验结果推断，因素 A 对指标有无显著影响，即当 A 取不同水平时指标有无显著差别。

A 取某个水平下的指标视为随机变量，判断 A 取不同水平时指标有无显著差别，相当于检验若干总体的均值是否相等。

1.1 数学模型

设 A 取 r 个水平 A_1, A_2, \dots, A_r ，在水平 A_i 下总体 x_i 服从正态分布 $N(\mu_i, \sigma^2)$ ， $i = 1, \dots, r$ ，这里 μ_i, σ^2 未知， μ_i 可以互不相同，但假定 x_i 有相同的方差。又设在每个水平 A_i 下作了 n_i 次独立试验，即从中抽取容量为 n_i 的样本，记作 $x_{ij}, j = 1, \dots, n_i$ ， x_{ij} 服从 $N(\mu_i, \sigma^2)$ ， $i = 1, \dots, r, j = 1, \dots, n_i$ 且相互独立。将这些数据列成表 1（单因素试验数据表）的形式。

表 1 单因素试验数据表

| | | | | |
|-------|----------|----------|-----|------------|
| A_1 | x_{11} | x_{12} | ... | x_{1n_1} |
| A_2 | x_{21} | x_{22} | ... | x_{2n_2} |
| ... | ... | ... | ... | ... |
| A_r | x_{r1} | x_{r2} | ... | x_{rn_r} |

将第 i 行称为第 i 组数据。判断 A 的 r 个水平对指标有无显著影响，相当于要作以下的假设检验

$$H_0: \mu_1 = \mu_2 = \dots = \mu_r; \quad H_1: \mu_1, \mu_2, \dots, \mu_r \text{ 不全相等}$$

由于 x_{ij} 的取值既受不同水平 A_i 的影响，又受 A_i 固定下随机因素的影响，所以将它分解为

$$x_{ij} = \mu_i + \varepsilon_{ij}, \quad i = 1, \dots, r, \quad j = 1, \dots, n_i \quad (1)$$

其中 $\varepsilon_{ij} \sim N(0, \sigma^2)$, 且相互独立。记

$$\mu = \frac{1}{n} \sum_{i=1}^r n_i \mu_i, \quad n = \sum_{i=1}^r n_i, \quad \alpha_i = \mu_i - \mu, \quad i = 1, \dots, r \quad (2)$$

μ 是总均值, α_i 是水平 A_i 对指标的效应。由 (1)、(2) 模型可表为

$$\begin{cases} x_{ij} = \mu + \alpha_i + \varepsilon_{ij} \\ \sum_{i=1}^r \alpha_i = 0 \\ \varepsilon_{ij} \sim N(0, \sigma^2), i = 1, \dots, r, j = 1, \dots, n_i \end{cases} \quad (3)$$

原假设为 (以后略去备选假设)

$$H_0: \alpha_1 = \alpha_2 = \dots = \alpha_r = 0 \quad (4)$$

1.2 统计分析

记

$$\bar{x}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}, \quad \bar{x} = \frac{1}{n} \sum_{i=1}^r \sum_{j=1}^{n_i} x_{ij} \quad (5)$$

$\bar{x}_{i\cdot}$ 是第 i 组数据的组平均值, \bar{x} 是总平均值。考察全体数据对 \bar{x} 的偏差平方和

$$S_T = \sum_{i=1}^r \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2 \quad (6)$$

经分解可得

$$S_T = \sum_{i=1}^r n_i (\bar{x}_{i\cdot} - \bar{x})^2 + \sum_{i=1}^r \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i\cdot})^2$$

记

$$S_A = \sum_{i=1}^r n_i (\bar{x}_{i\cdot} - \bar{x})^2 \quad (7)$$

$$S_E = \sum_{i=1}^r \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i\cdot})^2 \quad (8)$$

则

$$S_T = S_A + S_E \quad (9)$$

S_A 是各组均值对总方差的偏差平方和, 称为组间平方和; S_E 是各组内的数据对均值偏差平方和的总和。 S_A 反映 A 不同水平间的差异, S_E 则表示在同一水平下随机误差的大小。

注意到 $\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i\cdot})^2$ 是总体 $N(\mu_i, \sigma^2)$ 的样本方差的 $n_i - 1$ 倍, 于是有

$$\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i\cdot})^2 / \sigma^2 \sim \chi^2(n_i - 1)$$

由 χ^2 分布的可加性知

$$S_E/\sigma^2 \sim \chi^2\left(\sum_{i=1}^r (n_i - 1)\right)$$

即

$$S_E/\sigma^2 \sim \chi^2(n-r)$$

且有

$$ES_E = (n-r)\sigma^2 \quad (10)$$

对 S_A 作进一步分析可得

$$ES_A = (r-1)\sigma^2 + \sum_{i=1}^r n_i \alpha_i^2 \quad (11)$$

当 H_0 成立时

$$ES_A = (r-1)\sigma^2 \quad (12)$$

可知若 H_0 成立, S_A 只反映随机波动, 而若 H_0 不成立, 那它就还反映了 A 的不同水平的效应 α_i 。单从数值上看, 当 H_0 成立时, 由 (10)、(12) 对于一次试验应有

$$\frac{S_A/(r-1)}{S_E/(n-r)} \approx 1$$

而当 H_0 不成立时这个比值将远大于 1。当 H_0 成立时, 该比值服从自由度 $n_1 = r-1$, $n_2 = (n-r)$ 的 F 分布, 即

$$F = \frac{S_A/(r-1)}{S_E/(n-r)} \sim F(r-1, n-r) \quad (13)$$

为检验 H_0 , 给定显著性水平 α , 记 F 分布的 $1-\alpha$ 分位数为 $F_{1-\alpha}(r-1, (n-r))$, 检验规则为

$F < F_{1-\alpha}(r-1, (n-r))$ 时接受 H_0 , 否则拒绝。

以上对 S_A, S_E, S_T 的分析相当于对组间、组内等方差的分析, 所以这种假设检验方法称方差分析。

1.3 方差分析表

将试验数据按上述分析、计算的结果排成表 2 的形式, 称为单因素方差分析表 (Matlab 中给出的方差分析表)。

表 2 单因素方差分析表

| 方差来源 | 平方和 | 自由度 | 均方 | $1-p_r$ 分位数 | 概率 |
|--------|-------|-------|-------------------------------|-----------------------|-------|
| 因素 A | S_A | $r-1$ | $\bar{S}_A = \frac{S_A}{r-1}$ | $F_{1-p_r}(r-1, n-r)$ | p_r |
| 误差 | S_E | $n-r$ | $\bar{S}_E = \frac{S_E}{n-r}$ | | |
| 总和 | S_T | $n-1$ | | | |

最后一列给出大于 F 值的概率 p_r , $F_{1-p_r} < F_{1-\alpha}(r-1, (n-r))$ 相当于 $p_r > \alpha$ 。

方差分析一般用的显著性水平是：取 $\alpha = 0.01$ ，拒绝 H_0 ，称因素 A 的影响（或 A 各水平的差异）非常显著；取 $\alpha = 0.01$ ，不拒绝 H_0 ，但取 $\alpha = 0.05$ ，拒绝 H_0 ，称因素 A 的影响显著；取 $\alpha = 0.05$ ，不拒绝 H_0 ，称因素 A 无显著影响。

1.4 Matlab 实现

Matlab 统计工具箱中单因素方差分析的命令是 `anova1`。

若各组数据个数相等，称为均衡数据。若各组数据个数不等，称非均衡数据。

(1) 均衡数据

处理均衡数据的用法为：

`p=anova1(x)`

返回值 p 是一个概率，当 $p > \alpha$ 时接受 H_0 ， x 为 $m \times r$ 的数据矩阵， x 的每一列是一个水平的数据（这里各个水平上的样本容量 $n_i = m$ ）。另外，还输出一个方差表和一个 Box 图。

例 1 为考察 5 名工人的劳动生产率是否相同，记录了每人 4 天的产量，并算出其平均值，如表 3。你能从这些数据推断出他们的生产率有无显著差别吗？

表 3

| 工人 天 | A_1 | A_2 | A_3 | A_4 | A_5 |
|---------|-------|--------|--------|-------|-------|
| 1 | 256 | 254 | 250 | 248 | 236 |
| 2 | 242 | 330 | 277 | 280 | 252 |
| 3 | 280 | 290 | 230 | 305 | 220 |
| 4 | 298 | 295 | 302 | 289 | 252 |
| 平均产量 | 269 | 292.25 | 264.75 | 280.5 | 240 |

解 编写程序如下：

```
x=[ 256    254    250    248    236
    242    330    277    280    252
    280    290    230    305    220
    298    295    302    289    252];
```

```
p=anova1(x)
```

求得 $p = 0.1109 > \alpha = 0.05$ ，故接受 H_0 ，即 5 名工人的生产率没有显著差异。方差表对应于上面的单因素方差分析表的 1~4 列， $F = 2.262$ 是 $F(4,15)$ 分布的 $1-p$ 分位数，可以验证

`fcdf(2.262, 4, 15)=0.8891=1-p`

Box 图反映了各组数据的特征。

注：接受 H_0 ，是将 5 名工人的生产率作为一个整体进行假设检验的结果，并不表明取其中 2 个工人的生产率作两总体的均值检验时，也一定接受均值相等的假设。实际上，读者可以用 `ttest2` 对本题作 $H_0: \mu_2 = \mu_5$ 的检验，看看会得到什么结果。

(2) 非均衡数据

处理非均衡数据的用法为：

`p=anova1(x, group)`

x 为向量，从第 1 组到第 r 组数据依次排列；`group` 为与 x 同长度的向量，标志 x 中数据的组别（在与 x 第 i 组数据相对应的位置处输入整数 $i(i = 1, 2, \dots, r)$ ）。

例2 用4种工艺生产灯泡,从各种工艺制成的灯泡中各抽出了若干个测量其寿命,结果如下表,试推断这几种工艺制成的灯泡寿命是否有显著差异。

表4

| 工艺 序号 | A_1 | A_2 | A_3 | A_4 |
|----------|-------|-------|-------|-------|
| 1 | 1620 | 1580 | 1460 | 1500 |
| 2 | 1670 | 1600 | 1540 | 1550 |
| 3 | 1700 | 1640 | 1620 | 1610 |
| 4 | 1750 | 1720 | | 1680 |
| 5 | 1800 | | | |

解 编写程序如下:

```
x=[1620 1580 1460 1500
    1670 1600 1540 1550
    1700 1640 1620 1610
    1750 1720 1680 1800];
x=[x(1:4),x(16),x(5:8),x(9:11),x(12:15)];
g=[ones(1,5),2*ones(1,4),3*ones(1,3),4*ones(1,4)];
p=anova1(x,g)
```

求得 $0.01 < p = 0.0331 < 0.05$, 所以几种工艺制成的灯泡寿命有显著差异。

1.5 多重比较

在灯泡寿命问题中,为了确定哪几种工艺制成的灯泡寿命有显著差异,我们先算出各组数据的均值:

| 工艺 | A_1 | A_2 | A_3 | A_4 |
|----|-------|-------|-------|-------|
| 均值 | 1708 | 1635 | 1540 | 1585 |

虽然 A_1 的均值最大,但要判断它与其它几种有显著差异,还需做多重比较。一般多重比较要对所有 r 个总体作两两对比,分析相互间的差异。根据问题的具体情况可以减少对比次数。

对于上述问题, Matlab 多重比较的程序为

```
x=[1620 1580 1460 1500
    1670 1600 1540 1550
    1700 1640 1620 1610
    1750 1720 1680 1800];
x=[x(1:4),x(16),x(5:8),x(9:11),x(12:15)];
g=[ones(1,5),2*ones(1,4),3*ones(1,3),4*ones(1,4)];
[p,t,st]=anova1(x,g)
[c,m,h,nms] = multcompare(st);
[nms num2cell(m)]
```

§2 双因素方差分析

如果要考虑两个因素 A, B 对指标的影响, A, B 各划分几个水平,对每一个水平组合作若干次试验,对所得数据进行方差分析,检验两因素是否分别对指标有显著影响,或者还要进一步检验两因素是否对指标有显著的交互影响。

2.1 数学模型

设 A 取 r 个水平 A_1, A_2, \dots, A_r , B 取 s 个水平 B_1, B_2, \dots, B_s , 在水平组合 (A_i, B_j)

下总体 x_{ij} 服从正态分布 $N(\mu_{ij}, \sigma^2)$, $i=1, \dots, r$, $j=1, \dots, s$ 。又设在水平组合 (A_i, B_j) 下作了 t 个试验, 所得结果记作 x_{ijk} , x_{ijk} 服从 $N(\mu_{ij}, \sigma^2)$, $i=1, \dots, r$, $j=1, \dots, s$, $k=1, \dots, t$, 且相互独立。将这些数据列成表 5 的形式。

表 5 双因素试验数据表

| | B_1 | B_2 | \dots | B_s |
|----------|--------------------------|--------------------------|----------|--------------------------|
| A_1 | $x_{111} \cdots x_{11t}$ | $x_{121} \cdots x_{12t}$ | \dots | $x_{1s1} \cdots x_{1st}$ |
| A_2 | $x_{211} \cdots x_{21t}$ | $x_{221} \cdots x_{22t}$ | \dots | $x_{2s1} \cdots x_{2st}$ |
| \vdots | \vdots | \vdots | \vdots | \vdots |
| A_r | $x_{r11} \cdots x_{r1t}$ | $x_{r21} \cdots x_{r2t}$ | \dots | $x_{rs1} \cdots x_{rst}$ |

将 x_{ijk} 分解为

$$x_{ijk} = \mu_{ij} + \varepsilon_{ijk}, \quad i=1, \dots, r, \quad j=1, \dots, s, \quad k=1, \dots, t \quad (14)$$

其中 $\varepsilon_{ijk} \sim N(0, \sigma^2)$, 且相互独立。记

$$\begin{aligned} \mu &= \frac{1}{rs} \sum_{i=1}^r \sum_{j=1}^s \mu_{ij}, \quad \mu_{i\cdot} = \frac{1}{s} \sum_{j=1}^s \mu_{ij}, \quad \alpha_i = \mu_{i\cdot} - \mu \\ \mu_{\cdot j} &= \frac{1}{r} \sum_{i=1}^r \mu_{ij}, \quad \beta_j = \mu_{\cdot j} - \mu, \quad \gamma_{ij} = \mu_{ij} - \mu - \alpha_i - \beta_j \end{aligned} \quad (15)$$

μ 是总均值, α_i 是水平 A_i 对指标的效应, β_j 是水平 B_j 对指标的效应, γ_{ij} 是水平 A_i 与 B_j 对指标的交互效应。模型表为

$$\begin{cases} x_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} \\ \sum_{i=1}^r \alpha_i = 0, \quad \sum_{j=1}^s \beta_j = 0, \quad \sum_{i=1}^r \gamma_{ij} = \sum_{j=1}^s \gamma_{ij} = 0 \\ \varepsilon_{ijk} \sim N(0, \sigma^2), \quad i=1, \dots, r, \quad j=1, \dots, s, \quad k=1, \dots, t \end{cases} \quad (16)$$

原假设为

$$H_{01} : \alpha_i = 0 (i=1, \dots, r) \quad (17)$$

$$H_{02} : \beta_j = 0 (j=1, \dots, s) \quad (18)$$

$$H_{03} : \gamma_{ij} = 0 (i=1, \dots, r; j=1, \dots, s) \quad (19)$$

2.2 无交互影响的双因素方差分析

如果根据经验或某种分析能够事先判定两因素之间没有交互影响, 每组试验就不必重复, 即可令 $t=1$, 过程大为简化。

假设 $\gamma_{ij} = 0$, 于是

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

此时, 模型 (16) 可写成

$$\begin{cases} x_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \\ \sum_{i=1}^r \alpha_i = 0, \sum_{j=1}^s \beta_j = 0 \\ \varepsilon_{ij} \sim N(0, \sigma^2), i = 1, \dots, r, j = 1, \dots, s \end{cases} \quad (20)$$

对这个模型我们所要检验的假设为式 (17) 和式 (18)。下面采用与单因素方差分析模型类似的方法导出检验统计量。

记

$$\bar{x} = \frac{1}{rs} \sum_{i=1}^r \sum_{j=1}^s x_{ij}, \quad \bar{x}_{i\cdot} = \frac{1}{s} \sum_{j=1}^s x_{ij}, \quad \bar{x}_{\cdot j} = \frac{1}{r} \sum_{i=1}^r x_{ij}$$

$$S_T = \sum_{i=1}^r \sum_{j=1}^s (x_{ij} - \bar{x})^2$$

其中 S_T 为全部试验数据的总变差, 称为总平方和, 对其进行分解

$$\begin{aligned} S_T &= \sum_{i=1}^r \sum_{j=1}^s (x_{ij} - \bar{x})^2 \\ &= \sum_{i=1}^r \sum_{s=1}^s (x_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{x})^2 + s \sum_{i=1}^r (\bar{x}_{i\cdot} - \bar{x})^2 + r \sum_{j=1}^s (\bar{x}_{\cdot j} - \bar{x})^2 \\ &= S_E + S_A + S_B \end{aligned}$$

可以验证, 在上述平方和分解中交叉项均为 0。其中

$$S_E = \sum_{i=1}^r \sum_{s=1}^s (x_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{x})^2$$

$$S_A = s \sum_{i=1}^r (\bar{x}_{i\cdot} - \bar{x})^2, \quad S_B = r \sum_{j=1}^s (\bar{x}_{\cdot j} - \bar{x})^2$$

我们先来看看 S_A 的统计意义。因为 $\bar{x}_{i\cdot}$ 是水平 A_i 下所有观测值的平均, 所以

$\sum_{i=1}^r (\bar{x}_{i\cdot} - \bar{x})^2$ 反映了 $\bar{x}_{1\cdot}, \bar{x}_{2\cdot}, \dots, \bar{x}_{r\cdot}$ 差异的程度。这种差异是由于因素 A 的不同水平所

引起的, 因此 S_A 称为因素 A 的平方和。类似地, S_B 称为因素 B 的平方和。至于 S_E 的意义不甚明显, 我们可以这样来理解: 因为

$$S_E = S_T - S_A - S_B \quad (21)$$

在我们所考虑的两因素问题中, 除了因素 A 和 B 之外, 剩余的再没有其它系统性因素的影响, 因此从总平方和中减去 S_A 和 S_B 之后, 剩下的数据变差只能归入随机误差, 故 S_E 反映了试验的随机误差。

有了总平方和的分解式

$$S_T = S_E + S_A + S_B$$

以及各个平方和的统计意义, 我们就可以明白, 假设 (17) 的检验统计量应取为 S_A 与 S_E 的比。

和一元方差分析相类似, 可以证明, 当 H_{01} 成立时,

$$F_A = \frac{\frac{S_A}{r-1}}{\frac{S_E}{(r-1)(s-1)}} \sim F(r-1, (r-1)(s-1)) \quad (22)$$

当 H_{02} 成立时,

$$F_B = \frac{\frac{S_B}{s-1}}{\frac{S_E}{(r-1)(s-1)}} \sim F(s-1, (r-1)(s-1)) \quad (23)$$

检验规则为

$F_A < F_{1-\alpha}(r-1, (r-1)(s-1))$ 时接受 H_{01} , 否则拒绝 H_{01} ;

$F_B < F_{1-\alpha}(s-1, (r-1)(s-1))$ 时接受 H_{02} , 否则拒绝 H_{02} 。

我们可以写出方差分析表, 如表 6 所示。

表 6 无交互效应的两因素方差分析表

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

2.3 关于交互效应的双因素方差分析

与前面方法类似, 记

$$\bar{x} = \frac{1}{rst} \sum_{i=1}^r \sum_{j=1}^s \sum_{k=1}^t x_{ijk}, \quad \bar{x}_{ij\cdot} = \frac{1}{t} \sum_{k=1}^t x_{ijk}$$

$$\bar{x}_{i\cdot\cdot} = \frac{1}{st} \sum_{j=1}^s \sum_{k=1}^t x_{ijk}, \quad \bar{x}_{\cdot j\cdot} = \frac{1}{rt} \sum_{i=1}^r \sum_{k=1}^t x_{ijk}$$

将全体数据对 \bar{x} 的偏差平方和

$$S_T = \sum_{i=1}^r \sum_{j=1}^s \sum_{k=1}^t (x_{ijk} - \bar{x})^2 \quad (24)$$

进行分解, 可得

$$S_T = S_E + S_A + S_B + S_{AB} \quad (25)$$

其中

$$S_E = \sum_{i=1}^r \sum_{j=1}^s \sum_{k=1}^t (x_{ijk} - \bar{x}_{ij\cdot})^2, \quad S_A = st \sum_{i=1}^r (\bar{x}_{i\cdot\cdot} - \bar{x})^2$$

$$S_B = rt \sum_{j=1}^s (\bar{x}_{\cdot j \cdot} - \bar{x})^2, \quad S_{AB} = t \sum_{i=1}^r \sum_{j=1}^s (\bar{x}_{ij \cdot} - \bar{x}_{i \cdot \cdot} - \bar{x}_{\cdot j \cdot} + \bar{x})^2$$

称 S_E 为误差平方和, S_A 为因素 A 的平方和 (或行间平方和), S_B 为因素 B 的平方和 (或列间平方和), S_{AB} 为交互作用的平方和 (或格间平方和)。

可以证明, 当 H_{03} 成立时

$$F_{AB} = \frac{\frac{S_{AB}}{(r-1)(s-1)}}{\frac{S_E}{rs(t-1)}} \sim F((r-1)(s-1), rs(t-1)) \quad (26)$$

据此统计量, 可以检验 H_{03} 。

检验因子 A 和 B 的各个水平的效应是否有差异, 与 2.2 中的检验是一样的。

将试验数据按上述分析、计算的结果排成表 7 的形式, 称为双因素方差分析表。

表 7 关于交互效应的两因素方差分析表

| 方差来源 | 平方和 | 自由度 | 均方 | F 比 |
|------|----------|--------------|--|----------------------------------|
| 因素 A | S_A | $r-1$ | $\bar{S}_A = \frac{S_A}{r-1}$ | $\frac{\bar{S}_A}{\bar{S}_E}$ |
| 因素 B | S_B | $s-1$ | $\bar{S}_B = \frac{S_B}{s-1}$ | $\frac{\bar{S}_B}{\bar{S}_E}$ |
| 交互效应 | S_{AB} | $(r-1)(s-1)$ | $\bar{S}_{AB} = \frac{S_{AB}}{(r-1)(s-1)}$ | $\frac{\bar{S}_{AB}}{\bar{S}_E}$ |
| 误差 | S_E | $rs(t-1)$ | $\bar{S} = \frac{S_E}{rs(t-1)}$ | |
| 总和 | S_T | $rst-1$ | | |

2.4 Matlab 实现

统计工具箱中用 anova2 作双因素方差分析。命令为

p=anova2(x, reps)

其中 x 不同列的数据表示单一因素的变化情况, 不同行中的数据表示另一因素的变化情况。如果每种行—列对 (“单元”) 有不止一个的观测值, 则用参数 reps 来表明每个 “单元” 多个观测值的不同标号, 即 reps 给出重复试验的次数 t 。下面的矩阵中, 列因素有 3 种水平, 行因素有两种水平, 但每组水平有两组样本, 相应地用下标来标识:

$$\begin{bmatrix} x_{111} & x_{121} & x_{131} \\ x_{112} & x_{122} & x_{132} \\ x_{211} & x_{221} & x_{231} \\ x_{212} & x_{222} & x_{232} \end{bmatrix}$$

例 3 一种火箭使用了四种燃料、三种推进器, 进行射程试验, 对于每种燃料与每种推进器的组合作一次试验, 得到试验数据如表 8。问各种燃料之间及各种推进器之间有无显著差异?

表8 火箭试验数据

| | B_1 | B_2 | B_3 |
|-------|-------|-------|-------|
| A_1 | 58.2 | 56.2 | 65.3 |
| A_2 | 49.1 | 54.1 | 51.6 |
| A_3 | 60.1 | 70.9 | 39.2 |
| A_4 | 75.8 | 58.2 | 48.7 |

解 记燃料为因素 A ，它有 4 个水平，水平效应为 $\alpha_i, i=1,2,3,4$ 。推进器为因素 B ，它有 3 个水平，水平效应为 $\beta_j, j=1,2,3$ 。我们在显著性水平 $\alpha=0.05$ 下检验

$$H_1: \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$$

$$H_2: \beta_1 = \beta_2 = \beta_3 = 0$$

编写如下的 Matlab 程序：

```
x=[58.2 56.2 65.3
49.1 54.1 51.6
60.1 70.9 39.2
75.8 58.2 48.7];
[p,t,st]=anova2(x)
```

求得 $p=0.4491$ 0.7387，表明各种燃料和各种推进器之间的差异对于火箭射程无显著影响。

例 4 一火箭使用了 4 种燃料，3 种推进器作射程试验，每种燃料与每种推进器的组合各发射火箭 2 次，得到如表 9 结果。

表9 火箭试验数据

| | B_1 | B_2 | B_3 |
|-------|------------|------------|------------|
| A_1 | 58.2, 52.6 | 56.2, 41.2 | 65.3, 60.8 |
| A_2 | 49.1, 42.8 | 54.1, 50.5 | 51.6, 48.4 |
| A_3 | 60.1, 58.3 | 70.9, 73.2 | 39.2, 40.7 |
| A_4 | 75.8, 71.5 | 58.2, 51.0 | 48.7, 41.4 |

试在水平 0.05 下，检验不同燃料（因素 A ）、不同推进器（因素 B ）下的射程是否有显著差异？交互作用是否显著？

解 编写程序如下：

```
clc,clear
x0=[58.2,52.6 56.2,41.2 65.3,60.8
49.1,42.8 54.1,50.5 51.6,48.4
60.1,58.3 70.9,73.2 39.2,40.7
75.8,71.5 58.2,51.0 48.7,41.4];
x1=x0(:,1:2:5);x2=x0(:,2:2:6);
for i=1:4
x(2*i-1,:)=x1(i,:);
x(2*i,:)=x2(i,:);
end
[p,t,st]=anova2(x,2)
```

求得 $p=0.0035 \quad 0.0260 \quad 0.0001$ ，表明各试验均值相等的概率都为小概率，故可拒绝均值相等假设。即认为不同燃料（因素 A ）、不同推进器（因素 B ）下的射程有显著差异，交互作用也是显著的。

数据非均衡的双因素方差分析的 Matlab 命令要使用多因素方差分析的命令 `anovan`，具体使用方法参见下面的例 5。

§ 3 正交试验设计与方差分析

前面介绍了一个或两个因素的试验，由于因素较少，我们可以对不同因素的所有可能的水平组合做试验，这叫做全面试验。当因素较多时，虽然理论上仍可采用前面的方法进行全试验后再做相应的方差分析，但是在实际中有时会遇到试验次数太多的问题。如三因素四水平的问题，所有不同水平的组合有 $4^3=64$ 种，在每一种组合下只进行一次试验，也需做 64 次。如果考虑更多的因素及水平，则全试验的次数可能会大得惊人。因此在实际应用中，对于多因素做全试验是不现实的。于是我们考虑是否可以选择其中一部分组合进行试验，这就要用到试验设计方法选择合理的试验方案，使得试验次数不多，但也能得到比较满意的结果。

3.1 用正交表安排试验

正交表是一系列规格化的表格，每个表都有一个记号，如 $L_9(3^4)$ ，见表 10。

表 10 正交表 $L_9(3^4)$

| 列号 试验号 | 1 | 2 | 3 | 4 |
|-----------|---|---|---|---|
| 1 | 1 | 1 | 3 | 2 |
| 2 | 2 | 1 | 1 | 1 |
| 3 | 3 | 1 | 2 | 3 |
| 4 | 1 | 2 | 2 | 1 |
| 5 | 2 | 2 | 3 | 3 |
| 6 | 3 | 2 | 1 | 2 |
| 7 | 1 | 3 | 1 | 3 |
| 8 | 2 | 3 | 2 | 2 |
| 9 | 3 | 3 | 3 | 1 |

从表 10 可见， $L_9(3^4)$ 有 9 行，4 列，表中由数字 1, 2, 3 组成。

正交表的特点：

- (1) 每列中数字出现的次数相同，如 $L_9(3^4)$ 表每列中数字 1, 2, 3 均出现三次。
- (2) 任取两列数字的搭配是均衡的，如 $L_9(3^4)$ 表里每两列中 (1,1), (1,2), ..., (3,3)，九种组合各出现一次。

这种均衡性是一般正交表构造的特点，它使得根据正交表安排的试验，其试验结果具有很好的可比性，易于进行统计分析。

用正交表安排试验时，根据因素和水平个数的多少以及试验工作量的大小来考虑选用哪张正交表，下面举例说明。

例 5 为提高某种化学产品的转化率（%），考虑三个有关因素：反应温度 A （℃），反应时间 B （min）和使用催化剂的含量 C （%）。各因素选取三个水平，如表 11 所示。

表 11 转化率试验因素水平表

| 因素 水平 | 温度 <i>A</i> | 时间 <i>B</i> | 催化剂含量 <i>C</i> |
|----------|-------------|-------------|----------------|
| 1 | 80 | 90 | 5 |
| 2 | 85 | 120 | 6 |
| 3 | 90 | 150 | 7 |

如果做全面试验，则需 $3^3 = 27$ 次，若用正交表 $L_9(3^4)$ ，仅做 9 次试验。将三个因素 *A, B, C* 分别放在 $L_9(3^4)$ 表的任意三列上，如将 *A, B* 分别放在 $L_9(3^4)$ 的第 1, 2 列上，*C* 放在 $L_9(3^4)$ 的第 4 列上。将表中 *A, B, C* 所在的三列上的数字 1, 2, 3 分别用相应的因素水平去替代，得 9 次试验方案。以上工作称为表头设计。再将 9 次试验结果转化率数据列于表上（见表 12）。

表 12 转化率试验的正交表

| 因素 试验号 | 反应温度 <i>A</i> | 反应时间 <i>B</i> | 催化剂含量 <i>C</i> | 转化率 |
|-----------|---------------|---------------|----------------|-----|
| 1 | 80 (1) | 90 (1) | 6 (2) | 31 |
| 2 | 85 (2) | 90 (1) | 5 (1) | 54 |
| 3 | 90 (3) | 90 (1) | 7 (3) | 38 |
| 4 | 80 (1) | 120 (2) | 5 (1) | 53 |
| 5 | 85 (2) | 120 (2) | 7 (3) | 49 |
| 6 | 90 (3) | 120 (2) | 6 (2) | 42 |
| 7 | 80 (1) | 150 (3) | 7 (3) | 57 |
| 8 | 85 (2) | 150 (3) | 6 (2) | 62 |
| 9 | 90 (3) | 150 (3) | 5 (1) | 64 |

解 我们这里不作统计分析，直接利用 Matlab 多因素方差分析的函数 `anovan` 进行求解，程序如下：

```

y=[31 54 38 53 49 42 57 62 64];
g1=[1 2 3 1 2 3 1 2 3];
g2=[1 1 1 2 2 2 3 3 3];
g3=[2 1 3 1 3 2 3 2 1];
[p,t,st]=anovan(y,{g1,g2,g3})

```

求得概率 $p = 0.1364 \quad 0.0283 \quad 0.0714$ ，可见因素 *B, C* 的各水平对指标值的影响有显著差异（显著性水平取 0.1），而因素 *A* 的各水平对指标值的影响无显著差异。

习 题 十 一

1. 将抗生素注入人体会产生抗生素与血浆蛋白质结合的现象，以致减少了药效。表 13 列出 5 种常用的抗生素注入到牛的体内时，抗生素与血浆蛋白质结合的百分比。试在水平 $\alpha = 0.05$ 下检验这些百分比的均值有无显著的差异。设各总体服从正态分布，且方差相同。

表 13

| 青霉素 | 四环素 | 链霉素 | 红霉素 | 氯霉素 |
|------|------|------|------|------|
| 29.6 | 27.3 | 5.8 | 21.6 | 29.2 |
| 24.3 | 32.6 | 6.2 | 17.4 | 32.8 |
| 28.5 | 30.8 | 11.0 | 18.3 | 25.0 |
| 32.0 | 34.8 | 8.3 | 19.0 | 24.2 |

2. 为分析 4 种化肥和 3 个小麦品种对小麦产量的影响，把一块试验田等分成 36 小块，对种子和化肥的每一种组合种植 3 小块田，产量如表 14 所示（单位公斤），问品种、化肥及二者的交互作用对小麦产量有无显著影响。

表 14

| 化肥 | | A_1 | A_2 | A_3 | A_4 |
|----|-------|---------------|---------------|---------------|---------------|
| 品种 | B_1 | 173, 172, 173 | 174, 176, 178 | 177, 179, 176 | 172, 173, 174 |
| | B_2 | 175, 173, 176 | 178, 177, 179 | 174, 175, 173 | 170, 171, 172 |
| | B_3 | 177, 175, 176 | 174, 174, 175 | 174, 173, 174 | 169, 169, 170 |

3. (三因素方差分析) 某集团为了研究商品销售点所在的地理位置、销售点处的广告和销售点的装潢这三个因素对商品的影响程度，选了三个位置(如市中心黄金地段、非中心的地段、城乡结合部)，两种广告形式，两种装潢档次在四个城市进行了搭配试验。表15是销售量的数据，试在显著水平0.05下，检验不同地理位置、不同广告、不同装潢下的销售量是否有显著差异？

表15 三因素方差数据

| 城市号 水平组合 | 1 | 2 | 3 | 4 |
|-------------|-----|-----|-----|-----|
| $A_1B_1C_1$ | 955 | 967 | 960 | 980 |
| $A_1B_1C_2$ | 927 | 949 | 950 | 930 |
| $A_1B_2C_1$ | 905 | 930 | 910 | 920 |
| $A_1B_2C_2$ | 855 | 860 | 880 | 875 |
| $A_2B_1C_1$ | 880 | 890 | 895 | 900 |
| $A_2B_1C_2$ | 860 | 840 | 850 | 830 |
| $A_2B_2C_1$ | 870 | 865 | 850 | 860 |
| $A_2B_2C_2$ | 830 | 850 | 840 | 830 |
| $A_3B_1C_1$ | 875 | 888 | 900 | 892 |
| $A_3B_1C_2$ | 870 | 850 | 847 | 965 |
| $A_3B_2C_1$ | 870 | 863 | 845 | 855 |
| $A_3B_2C_2$ | 821 | 842 | 832 | 848 |