方差分析

1　单因素方差分析

2　双因素方差分析

- 无交互作用的双因素方差分析
- 有交互作用的双因素方差分析

方差分析

实际问题中影响某个（或某些）指标的**因素**有很多，而这些因素由于试验条件的限制，往往只可以取有限种状态或只能定性地描述(称为**水平**，取**离散值**)，要找出在众多因素中找出有**显著影响**的因素。

例如：影响农作物的单位面积产量有品种、施肥种类、施肥量等许多因素，需要找出对农作物的单位面积产量有显著影响的因素。

为解决这类问题，需要**先做试验，然后对实验结果进行分析，作出判断。**

(1)试验设计: 要求试验方案能很好地反映观测因素的影响作用，且试验次数尽可能少，以节约人力、物力和时间；

(2)方差分析: 通过观测数据**推断某个因素的影响是否显著**，这类统计方法称为方差分析(ANalysis Of Variance, ANOVA).

所关心的试验结果称为试验**指标**(因变量);

试验中可以变化的、影响试验指标的因素称为**因素(因子)**(自变量),用大写字母A、B、C、⋯表示;

因素在试验中所取的不同状态或等级称为**水平**,**因素$\boldsymbol{A}$的** r 个不同水平用$A_1,\cdots,A_r$表示.

- **单因素方差分析:**只考虑一个因素的不同水平的影响
- **双因素方差分析:**考虑两个因素的不同水平的影响
- **多因素方差分析:**考虑两个以上因素的不同水平的影响

例 三名实验员各用四种不同型号的仪器对同一物理常数进行测定,希望了解不同实验员及不同型号的仪器对测定值有何影响.

这时,实验员和测量仪器是因素,分别记为 A 和 B,

因素 A 有 3 个水平,记为 A_1,A_2,A_3;

因素 B 有 4 个水平,记为 B_1,B_2,B_3,B_4.

这是双因素方差分析问题。

1 单因素方差分析

例 某灯泡厂用四种不同的灯丝生产灯泡. 从每种灯泡中随机抽取若干个灯泡测其寿命(单位:小时), 得数据如表. 问这四种灯丝生产的灯泡的使用寿命是否有显著差异?

	1	2	3	4	5	6	7	8	平均值
甲	1600	1610	1650	1680	1700	1700	1780		**1674**
乙	1500	1640	1400	1700	1750				**1598**
丙	1640	1550	1600	1620	1640	1600	1740	1800	**1649**
丁	1510	1640	1530	1570	1520	1680			**1575**

这里试验**指标**为灯泡的使用寿命, **因素**为灯丝, 有四个**水平**。

从表中可以看出

- 采用同一种灯丝生产的灯泡,其使用寿命也有差异,说明使用寿命是随机变量;
- 采用不同灯丝生产的灯泡, 其使用寿命的均值有一定差异.

本例中仅考虑灯丝这一因素对灯泡寿命的影响，上述数据不能认为出自同一个总体，而是把每个水平下所关心的因变量看作一个总体，这里认为同一种灯丝生产的灯泡就是一个总体，分别从四个总体$X^{(1)}, X^{(2)}, X^{(3)}, X^{(4)}$抽取容量分别为 7，5，8，6 的样本观测值.

在方差分析中假定各总体相互独立，且服从同方差的正态分布，即第 i 种灯丝生产的灯泡寿命$X^{(i)}$是一随机变量

$$X^{(i)} \sim N(\mu_i, \sigma^2) \quad, i = 1,2,3,4$$

- 随机因素既可能造成**同一总体内各样本的差异(组内差异)**，也可能造成**不同总体各样本的差异(组间差异)**.
  - 若灯丝对灯泡寿命的影响**不显著**，则上述两种差异都主要来源于随机因素的影响，二者的差别不会很大；
  - 若灯丝对灯泡寿命的影响是**显著**的，则**组间差异**主要来源于可控因素的影响，应该远大于**组内差异**。
- 可以通过比较这两种差异的大小来检验因素（灯丝）是否对试验结果（寿命）有显著影响。

试验的目的就是检验假设:

$$H_0: \mu_1 = \cdots = \mu_4, \qquad H_1: \mu_1, \cdots, \mu_4\text{不全相同}$$

若原假设H_0成立,则认为不同灯丝的灯泡寿命X_i没有显著差异，即灯泡寿命差异只是由其它随机因素引起的.

- 数据的差异经常用**方差**来描述，这种考察均值之间差异的方法称为“**方差分析**”.
- **方差分析**是将所有样本表现出来的差异（方差）按照来源分解为多个组成部分, 然后进行比较, 用以检验某种因素是否对试验结果产生显著的影响. 总体来看, 方差分析的核心就是**变异分解**.

问题：方差分析是比较各组的均值是否有差异，*t*检验也可以用于比较两组均值是否有差异，可否将*t*检验应用于方差分析的问题?

方差分析不能用t检验。t检验只能用于比较两组的均值是否有差异，采用t检验检验多组均值需要进行两两比较，这样犯第一类错误的概率就会增大。

例如：设$\alpha=0.05$，则两组均值进行检验时，犯第一类错误的概率为0.05；但若是进行三组均值的两两比较，则需要进行三次t检验，此时犯错误的概率为$1-(1-0.05)^3=0.1426$，该值远大于0.05.

多次检验的错误率

假设进行 n 次独立的假设检验，每次检验 I 类错误率是 α.

$P(\text{at least 1 Type I Error}) = 1 - P(\text{no Type I Errors}) = 1 - (1-\alpha)^n$

例如 $n = 20$，$\alpha = 0.05$, $1-(0.95)^{20} \approx 64.2\%$

p值篡改(p-hacking)指研究人员每次使用不同的数据，不断的统计直到统计结果满意。

例如，在统计时根据p值进行取舍，统计分析后发现结果不满意，然后删除一些数据再统计，直到结果满意。

p-hacking有太多的人为因素参与，很易导致结果的不准确性，也就是假阳性。

这也会引起实验结果的不可重复性, 另外一个实验室甚至研究者本人再做一遍也得不到同样的结果。

数学模型

一般地设因素A有r个不同水平$A_1,\ldots,A_r$,在A_i下试验结果$X^{(i)} \sim N(\mu_i,\sigma^2)$, $i=1,\ldots,r$. 在A_i下做了$n_i(\geq 2)$次试验,相当于从总体$X^{(i)}$中抽取一组样本$X_{i1},\ldots,X_{in_i}$, 相互独立, 方差分析模型为:

$$X_{ij}=\mu_i+\varepsilon_{ij} \qquad \varepsilon_{ij} \sim N(0;\sigma^2) \qquad \begin{cases} X_{ij} \sim N(\mu_i,\sigma^2) \\ X_{ij} \text{ 相互独立} \quad ,i=1,2,\cdots,r \quad ;j=1,2,\cdots,n_i \\ \mu_i,\sigma^2 \text{ 未知} \end{cases}$$

ε_{ij}表示随机因素对A_i下第j个指标的影响,称为随机误差.

检验问题: $H_0:\mu_1=\cdots=\mu_r \leftrightarrow H_1:\mu_1,\cdots,\mu_r$不全相同

拒绝H_0表示因素A的作用显著, 否则认为因素A不显著.

对模型作等价变形

$$\varepsilon_{ij} = X_{ij} - \mu_i,\quad n = \sum_{i=1}^{r} n_i\,,\quad \mu = \frac{1}{n}\sum_{i=1}^{r} n_i \cdot \mu_i$$ 称为理论总值;

$\alpha_i = \mu_i - \mu$, $(i = 1, \cdots, r; j = 1, \cdots, n_i)$ 称为该因素的第i水平A_i 的效应

等价方差分析模型为

$$\begin{cases} X_{ij} = \mu + \alpha_i + \varepsilon_{ij}, i = 1, \cdots, r; j = 1, \cdots, n_i \\ \varepsilon_{ij} \sim N(0, \sigma^2) \text{ 且相互独立} \\ \sum_{i=1}^{r} n_i \cdot \alpha_i = 0 \\ \mu, \alpha_i, \sigma^2 \text{ 未知} \end{cases}$$

等价检验问题为 $H_0\!:\alpha_1 = \cdots = \alpha_r = 0$，$H_1\!:\alpha_1, \cdots, \alpha_r$不全为0

试验指标数据可列成下表形式

因素水平	总体	样本观测数据
A_1	$X^{(1)} \sim N(\mu+\alpha_1, \sigma^2)$	$X_{11}, X_{12}, \dots, X_{1n_1}$
A_2	$X^{(2)} \sim N(\mu+\alpha_2, \sigma^2)$	$X_{21}, X_{22}, \dots, X_{2n_2}$
$\vdots$	$\vdots$	$\vdots$
A_r	$X^{(r)} \sim N(\mu+\alpha_r, \sigma^2)$	$X_{r1}, X_{r2}, \dots, X_{rn_r}$

平方和分解公式

记 $X_i=\sum_{j=1}^{n_i}X_{ij}$ ， $\overline{X}_i=\frac{1}{n_i}\sum_{j=1}^{n_i}X_{ij}$ ， $\overline{X}=\frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{n_i}X_{ij}$, $n=\sum_{i=1}^{r}n_i$

$S_T=\sum_{i=1}^{r}\sum_{j=1}^{n_i}(X_{ij}-\overline{X})^2=\sum_{i=1}^{r}\sum_{j=1}^{n_i}[(X_{ij}-\overline{X}_i)+(\overline{X}_i-\overline{X})]^2$

$$=\sum_{i=1}^{r}\sum_{j=1}^{n_i}(X_{ij}-\overline{X}_i)^2+\sum_{i=1}^{r}\sum_{j=1}^{n_i}(\overline{X}_i-\overline{X})^2+2\sum_{i=1}^{r}\sum_{j=1}^{n_i}(X_{ij}-\overline{X}_i)(\overline{X}_i-\overline{X})$$

$$=\sum_{i=1}^{r}\sum_{j=1}^{n_i}(X_{ij}-\overline{X}_i)^2+\sum_{i=1}^{r}\sum_{j=1}^{n_i}(\overline{X}_i-\overline{X})^2$$

$=\sum_{i=1}^{r}\sum_{j=1}^{n_i}(X_{ij}-\overline{X}_i)^2+\sum_{i=1}^{r}n_i\cdot(\overline{X}_i-\overline{X})^2=S_e+S_A$

其中 $\sum_{i=1}^{r}\sum_{j=1}^{n_i}(X_{ij}-\overline{X}_i)(\overline{X}_i-\overline{X})=\sum_{i=1}^{r}(\overline{X}_i-\overline{X})\sum_{j=1}^{n_i}(X_{ij}-\overline{X}_i)$

$=\sum_{i=1}^{r}(\overline{X}_i-\overline{X})(\sum_{j=1}^{n_i}X_{ij}-n_i\cdot\overline{X}_i)=\sum_{i=1}^{r}(\overline{X}_i-\overline{X})(n_i\cdot\overline{X}_i-n_i\cdot\overline{X}_i)=0$

平方和分解公式 $S_T = S_e + S_A$

$$\sum_{i=1}^{r}\sum_{j=1}^{n_i}(X_{ij}-\overline{X})^2 = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(X_{ij}-\overline{X}_i)^2 + \sum_{i=1}^{r} n_i \cdot (\overline{X}_i-\overline{X})^2$$

- $S_T = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(X_{ij}-\overline{X})^2$ **总偏差平方和**
- $S_e = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(X_{ij}-\overline{X}_i)^2$ **误差平方和(组内平方和)**

  完全由随机因素决定。
- $S_A = \sum_{i=1}^{r} n_i \cdot (\overline{X}_i-\overline{X})^2$ **因素平方和(组间平方和)**,
  - 若假设H_0成立, S_A完全由随机因素决定;
  - 若假设H_0不成立, S_A除反映随机因素引起的波动之外, 还反映了因素不同水平效应的差异所引起的波动.

定理 在单因素方差分析问题中，

（1）S_A 与 S_e 相互独立

（2）S_e/σ^2 服从 $\chi^2(n-r)$， $E(S_e)=(n-r)\sigma^2$

（3）$E(S_A)=(r-1)\sigma^2+\sum_{i=1}^{r} n_i\alpha_i^2$

当 $H_0\colon \alpha_1=\cdots=\alpha_r=0$ 成立时，S_A/σ^2 服从 $\chi^2(r-1)$

证 (1) 对每一个$i(i=1,2,\cdots,r)$, 总体$X^{(i)}$的样本均值$\bar{X}_i$与样本方差$\frac{1}{n_i-1}\sum_{j=1}^{n_i}\left(X_{ij}-\bar{X}_i\right)^2$ 相互独立; 又由全体样本相互独立知

随机变量 $\bar{X}_1,\cdots,\bar{X}_r,\sum_{j=1}^{n_1}\left(X_{1j}-\bar{X}_1\right)^2,\cdots,\sum_{j=1}^{n_r}\left(X_{rj}-\bar{X}_r\right)^2$ 相互独立.

从而 $\left(\bar{X}_1,\bar{X}_2,\cdots,\bar{X}_r\right)$ 与 $S_e=\sum_{i=1}^{r}\sum_{j=1}^{n_i}\left(X_{ij}-\bar{X}_i\right)^2$ 相互独立, 由此

$S_A=\sum_{i=1}^{r}n_i\cdot(\overline{X}_i-\overline{X})^2$ 与 S_e 相互独立.

(2) 对于第i个总体, 有 $\sum_{j=1}^{n_i}\left(X_{ij}-\bar{X}_i\right)^2/\sigma^2\sim\chi^2(n_i-1),i=1,2,\cdots,r,$

因而得到

$$\frac{S_e}{\sigma^2}=\sum_{i=1}^{r}\frac{\sum_{j=1}^{n_i}\left(X_{ij}-\bar{X}_i\right)^2}{\sigma^2}\sim\chi^2(n-r)$$

（3）由于 $\bar{X}_i \sim N\left(\mu+\alpha_i, \frac{\sigma^2}{n_i}\right)$, 且 $\bar{X}_1, \bar{X}_2, \cdots, \bar{X}_r$ 相互独立, 所以

$$\bar{X}=\frac{1}{n}\sum_{i=1}^{r} n_i \bar{X}_i \sim N\left(\mu, \frac{\sigma^2}{n}\right), \quad cov(\bar{X}_i, \bar{X})=\frac{\sigma^2}{n}$$

$$E(\bar{X}_i-\bar{X})=\alpha_i,\ D(\bar{X}_i-\bar{X})=\frac{n-n_i}{nn_i}\sigma^2$$

$$E(\bar{X}_i-\bar{X})^2=\alpha_i^2+\frac{n-n_i}{nn_i}\sigma^2$$

$$E(S_A)=E\left[\sum_{i=1}^{r} n_i(\bar{X}_i-\bar{X})^2\right]=\sum_{i=1}^{r} n_i\left(\alpha_i^2+\frac{n-n_i}{nn_i}\sigma^2\right)$$

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

$$E\left(\frac{S_e}{n-r}\right)=\sigma^2,\ E\left(\frac{S_A}{r-1}\right)=\sigma^2+\frac{1}{r-1}\sum_{i=1}^{r} n_i\alpha_i^2.$$

- 当 H_0 成立时，统计量 $S_e/(n-r)$ 与 $S_A/(r-1)$ 都是 σ^2 的无偏估计，比值 $\frac{S_A/(r-1)}{S_e/(n-r)}$ 应接近于 1；

- 当 H_0 不成立时，有 $E\left(\frac{S_A}{r-1}\right)>\sigma^2$，比值 $\frac{S_A/(r-1)}{S_e/(n-r)}$ 应比 1 明显的偏大

统计量 $F=\frac{S_A/(r-1)}{S_e/(n-r)}$ 可作为 H_0 的检验统计量.

检验统计量 $$F=\frac{S_A/(r-1)}{S_e/(n-r)} \overset{\text{若}H_0\text{成立}}{\sim} F(r-1,n-r)$$

检验水平a下的拒绝域

$$W=\{(x_1,\dots,x_n): F>F_\alpha(r-1,n-r)\}$$

若$F>F_\alpha(r\text{-}1,n\text{-}r)$，认为因素取不同水平对试验指标有显著影响.

针对不同的检验水平a取值情况:

$F > F_{0.01}(r\text{-}1,n\text{-}r)$ 认为因素A的影响**高度显著**,用**表示

$F_{0.05} < F \leq F_{0.01}(r\text{-}1,n\text{-}r)$ 认为因素A的影响**显著**,用*表示

$F_{0.1} < F \leq F_{0.05}(r\text{-}1,n\text{-}r)$ 认为因素A有**一定影响**,用(*)表示

$F \leq F_{0.1}(r\text{-}1,n\text{-}r)$ 认为因素A的影响**不显著**

说明：利用方差分析检验因素取不同水平对试验结果的影响是否显著，是相对随机因素而言的"显著". 如果随机因素的影响很小，即使判断出因素不同水平的影响具有 "显著性"，其绝对意义下的影响也可能较小.

单因素方差分析表

来源	平方和	自由度	均方和	F 比	临界值	显著性
（组间）因素 A	S_A	$r-1$	$S_A/(r-1)$	$F=\dfrac{S_A/(r-1)}{S_e/(n-r)}$	$F_\alpha(r-1,n-r)$	
（组内）误差 e	$S_e=S_T-S_A$	$n-r$	$S_e/(n-r)$			
总　和	S_T	$n-1$				

计算S_T、S_A和 S_e 的公式:

$$S_T=\sum_{i=1}^{r}\sum_{j=1}^{n_i}{X_{ij}}^2-n\,(\bar{X})^2$$

$$S_A=\sum_{i=1}^{r}n_i(\bar{X}_i)^2-\ n\cdot(\bar{X})^2$$

$$S_e=S_T-S_A$$

例 某灯泡厂用四种不同的灯丝生产四种灯泡. 从每种灯泡中随机抽取若干个灯泡测其寿命(单位:小时), 得如下数据. 试问这四种灯丝生产的灯泡的使用寿命是否有显著差异？ (α=0.05)

	1	2	3	4	5	6	7	8	平均值
甲	1600	1610	1650	1680	1700	1700	1780		**1674**
乙	1500	1640	1400	1700	1750				**1598**
丙	1640	1550	1600	1620	1640	1600	1740	1800	**1649**
丁	1510	1640	1530	1570	1520	1680			**1575**

这里试验**指标**为灯泡的使用寿命, **因素**为灯丝, 有四个**水平**

$r=4,\ n_1=7, n_2=5, n_3=8, n_4=6,\ n = n_1+n_2+n_3+n_4=26$

方差分析表

来源	平方和	自由度	均方和	F 比	临界值	显著性
因素A	39776.4	3	13258.3	1.638	$F_{0.1}=$ 2.35	
误差e	178089	22	8095			
总和	217865.4	25				

由于 $F=1.638 < 2.35 = F_{0.1}(3, 22)$，故接受 H_0，即认为灯丝对灯泡的寿命没有显著影响.

```
from scipy.stats import f_oneway
group_A = np.array([1600,1610,1650,1680,1700,1700,1780])
group_B = np.array([1500,1640, 1400, 1700, 1750])
group_C = np.array([1640, 1550, 1600, 1620, 1640, 1600, 1740, 1800])
group_D = np.array([1510, 1520, 1530, 1570, 1640, 1680]);
# One-way ANOVA
f_stat, p_val = f_oneway(group_A, group_B, group_C, group_D)
print(f"ANOVA test: F={f_stat}, p={p_val}")
```

例 考查燃烧温度对砖的密度的影响, 观测 4 种燃烧温度下砖的密度, 得如下数据

	1	2	3	4	5
100℃	21.8	21.9	21.7	21.6	21.7
125℃	21.7	21.4	21.5	21.4	
150℃	22.9	22.8	22.8	22.6	22.5
175℃	21.9	21.7	21.8	21.4	

问燃烧温度对砖密度的影响是否显著？(α=0.01)

解 $r=4,\ n_1=5, n_2=4, n_3=5, n_4=4$

$n = n_1+n_2+n_3+n_4=18$

方差分析表

来源	平方和	自由度	均方和	F 比	临界值	显著性
因素A	424.5	3	141.5	55.03	$F_{0.01}=$ 5.56	**
误差e	36	14	2.571			
总和	460.5	17				

由于 $F=55.03>5.56=F_{0.01}(3,14)$，故拒绝 H_0，

即认为温度对砖密度的影响高度显著

2 双因素方差分析

设因素 A 取 a 个不同的水平 $A_1, A_2, \cdots, A_a$, 因素B 取 b 个不同的水平 $B_1, B_2, \cdots, B_b$. 把因素A 取水平 A_i 且因素B 取水平 B_j 时, 所关心的因变量看作一个总体 $X^{(ij)}$, 并设 $X^{(ij)} \sim N\left(\mu_{ij}; \sigma^2\right)$.

从每个总体 $X^{(ij)}$ 抽取一个容量为 m 的样本 $\left(X_{ij1}, X_{ij2}, \cdots, X_{ijm}\right), i = 1,2, \cdots, a, j = 1,2, \cdots, b$, 并且假定所有 X_{ijk} 相互独立.

建立下面的正态线性模型来讨论双因素方差分析问题:

$$X_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk},$$

$$k = 1,2, \cdots, m;\ i = 1,2, \cdots, a, j = 1,2, \cdots, b$$

其中 $\varepsilon_{111}, \varepsilon_{112}, \cdots, \varepsilon_{11m}; \cdots; \varepsilon_{ab1}, \varepsilon_{db2}, \cdots, \varepsilon_{abm}$ 是独立同分布的随机变量, 且每一个 $\varepsilon_{ijk} \sim N\left(0; \sigma^2\right)$

$$总平均\ \mu = \frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\mu_{ij},$$

$$\mu_{i.} = \frac{1}{b}\sum_{j=1}^{b}\mu_{ij},\ \alpha_i = \mu_{i.} - \mu, i = 1,2,\cdots,a,$$

$$\mu_{.j} = \frac{1}{a}\sum_{i=1}^{a}\mu_{ij},\ \beta_j = \mu_{.j} - \mu, j = 1,2,\cdots,b.$$

α_i 称为因素A 在第 i 个水平 A_i 下的**效应**, β_j 称为因素B 在第 j 个水 平 B_j 下的**效应**. 显然 $\sum_{i=1}^{a}\alpha_i = \sum_{j=1}^{b}\beta_j = 0.$

$$\gamma_{ij} = \mu_{ij} - (\mu + \alpha_i + \beta_j) = \mu_{ij} - \mu_{i.} - \mu_{.j} + \mu,\quad i = 1,2,\cdots,a,\ j = 1,2,\cdots,b,$$

$$\begin{cases} \sum_{j=1}^{b}\gamma_{ij} = 0,\ i = 1,2,\cdots,a, \\ \sum_{i=1}^{a}\gamma_{ij} = 0,\ j = 1,2,\cdots,b, \end{cases}$$

γ_{ij} 称为因素A 的第 i 个水平 A_i 与因素B 的第 j 个水平 B_j 之间的**交互效应**.

- 如果 $\gamma_{ij} = 0$，为**无交互作用的双因素方差分析问题**.这时两个因素的联合作用可以由每个因素效应的简单迭加来表示.

2.1 无交互作用的双因素方差分析

$$X_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, i = 1,2,\cdots,a, j = 1,2,\cdots,b;$$

$$\sum_{i=1}^{a} \alpha_i = \sum_{j=1}^{b} \beta_j = 0;$$

$\varepsilon_{11}, \varepsilon_{12}, \cdots, \varepsilon_{ab}$ 是独立同分布的随机变量,且每一个 $\varepsilon_{ij} \sim N\left(0;\sigma^2\right)$

对两个因素A 和 B的每一对水平搭配 $\left(A_i, B_j\right)$, 有容量为 1 的样本.通常称其为**双因素每对水平搭配只观测一次(或非重复)试验的方差分析**.

$$\bar{X}_{i\cdot} = \frac{1}{b}\sum_{j=1}^{b} X_{ij}, i = 1,2,\cdots,a,$$

$$\bar{X}_{\cdot j} = \frac{1}{a}\sum_{i=1}^{a} X_{ij}, j = 1,2,\cdots,b,$$

$$\bar{X} = \frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b} X_{ij} = \frac{1}{a}\sum_{i=1}^{a} \bar{X}_{i\cdot} = \frac{1}{b}\sum_{j=1}^{b} \bar{X}_{\cdot j} \qquad n = ab$$

平方和分解公式 在无交互作用的双因素方差分析模型中，

$$SS = SS_A + SS_B + SS_e.$$

- $SS = \sum_{i=1}^{a}\sum_{j=1}^{b}\left(X_{ij} - \bar{X}\right)^2$ 总离差平方和，反映全体数据中的波动
- $SS_A = b\sum_{i=1}^{a}(\bar{X}_{i\cdot} - \bar{X})^2$ 由因素 A 引起的离差平方和，反映由于因素A 在各个水平下的不同作用而引起的波动
- $SS_B = a\sum_{j=1}^{b}\left(\bar{X}_{\cdot j} - \bar{X}\right)^2$ 由因素 B 引起的离差平方和，反映由于因素B 在各个水平下的不同作用而引起的波动
- $SS_e = \sum_{i=1}^{a}\sum_{j=1}^{b}\left(X_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X}\right)^2$ 误差平方和，反映由于随机误差的作用而在数据中引起的波动

要分别检验假设

$\mathrm{H}_{0A}: \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$ 及 $\mathrm{H}_{0B}: \beta_1 = \beta_2 = \cdots = \beta_b = 0$.

定理 在无交互作用的双因素方差分析模型中,

(1) $\mathrm{SS}_A, \mathrm{SS}_B, \mathrm{SS}_e$ 相互独立

(2) $SS_e/\sigma^2 \sim \chi^2(n-a-b+1)$.

(3) 当 H_{0A} 成立, 即 $\alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$ 时, $SS_A/\sigma^2 \sim \chi^2(a-1)$;

(4) 当 H_{0B} 成立, 即 $\beta_1 = \beta_2 = \cdots = \beta_b = 0$ 时, $SS_B/\sigma^2 \sim \chi^2(b-1)$.

假设检验$H_{0A}: \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$

取检验统计量为

$$F_A = \frac{SS_A/(a-1)}{SS_e/(n-a-b+1)},$$

- 当 H_{0A} 成立时, $F_A \sim F(a-1, n-a-b+1)$,
- 当 H_{0A} 不成立时, 由于

$$E(SS_A) = bE\left[\sum_{i=1}^{a} (\bar{X}_{i\cdot} - \bar{X})^2\right] = b\sum_{i=1}^{a} E[(\bar{X}_{i\cdot} - \bar{X})^2] = (a-1)\sigma^2 + b\sum_{i=1}^{a} \alpha_i^2,$$

所以 SS_A 有偏大的趋势, 从而由

$$F_A > F_\alpha(a-1, n-a-b+1)$$

所确定的拒绝域给出了显著性水平 α 下的一个检验.

假设检验H_{0B}: $\boldsymbol{\beta_1 = \beta_2 = \cdots = \beta_b = 0}$

取检验统计量为

$$F_B = \frac{SS_B/(b-1)}{SS_e/(n-a-b+1)},$$

- 当 H_{0B} 成立, 即 $\beta_1 = \beta_2 = \cdots = \beta_b = 0$ 时, $F_B \sim F(b-1, n-a-b+1)$;
- 当 H_{0B} 不成立时， 由于

$$E(SS_B) = aE\left[\sum_{j=1}^{b} \left(\bar{X}_{.j} - \bar{X}\right)^2\right] = a\sum_{j=1}^{b} E\left[\left(\bar{X}_{.j} - \bar{X}\right)^2\right] = (b-1)\sigma^2 + a\sum_{j=1}^{b} \beta_j^2,$$

所以 SS_B 有偏大的趋势, 从而由

$$F_B > F_\alpha(b-1, n-a-b+1)$$

所确定的拒绝域给出了显著性水平 α 下的一个检验.

无交互作用的双因素方差分析表

方差来源	平方和	自由度	均方和	F 值
因素A	$SS_A = b\sum_{i=1}^{a}(\bar{X}_{i\cdot} - \bar{X})^2$	$a-1$	$MS_A = \dfrac{SS_A}{a-1}$	$F_A = \dfrac{MS_A}{MS_e}$
因素B	$SS_B = a\sum_{j=1}^{b}\left(\bar{X}_{\cdot j} - \bar{X}\right)^2$	$b-1$	$MS_B = \dfrac{SS_B}{b-1}$	$F_B = \dfrac{MS_B}{MS_e}$
误差	$SS_e = \sum_{i=1}^{a}\sum_{j=1}^{b}\left(X_{ij} - \bar{X}_{i\cdot} - \bar{X}\bar{X}_{\cdot j} + \bar{X}\right)^2$	$n-a-b+1$	$MS_e = \dfrac{S_e}{n-a-b+1}$	
总和	$SS = \sum_{i=1}^{a}\sum_{j=1}^{b}\left(X_{ij} - \bar{X}\right)^2$	$n-1$		

例 某型号火箭使用了四种燃料、三种推进器做射程试验. 每种燃料与每种推进器搭配做一次试验, 测得的火箭射程(单位: km) 如下, 在显著性水平 α=0.05 下, 燃料与推进器对射程是否有显著影响?

射程 推进器 B / 燃料 A	B_1	B_2	B_3	$\bar{x}_{i\cdot}$
A_1	158.2	156.2	165.3	159.90
A_2	149.1	154.1	151.6	151.60
A_3	160.1	170.9	139.2	156.73
A_4	175.8	158.2	148.7	160.90
$\bar{x}_{\cdot j}$	160.80	159.88	151.20	$\bar{x}$=157.28

解 由所给数据列出方差分析表 $(a=4, b=3, n=12)$：

方差来源	平方和	自由度	均方和	F 值
因素A	157.59	3	52.53	0.43
因素B	223.85	2	111.93	0.92
误差	731.98	6	**121.99**	
总和	1113.42	11		

- 对于假设检验 H_{0A}, 查表得临界值 $F_{0.05}(3,6)=4.76>0.43$, 因此不能拒绝 H_{0A}, 即各种燃料的差异对火箭射程的影响不显著.
- 对于假设检验 H_{0B}, 查表得临界值 $F_{0.05}(2,6)=5.14>0.92$, 因此也不能拒绝 H_{0B}, 即各种推进器的差异对火箭射程的影响并不显著.

本例中误差均方和 MS_e出现较大的值, 这可能是没有考虑因素 A 与 B 的不同水平搭配所引起的交互作用的缘故.

2.2 有交互作用的双因素方差分析

$$X_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk},$$

$$i = 1,2,\cdots,a, j = 1,2,\cdots,b, k = 1,2,\cdots,m$$

有两个因素 A 和 B, 因素A 取 a 个不同水平, 因素B 取 b 个不同水平,

对这两个因素的每一对水平搭配(A_i, B_j), 有容量为$m(>1)$的样本.

$$\bar{X}_{ij\cdot} = \frac{1}{m}\sum_{k=1}^{m} X_{ijk}, i = 1,2,\cdots,a, j = 1,2,\cdots,b;$$

$$\bar{X}_{i\cdot\cdot} = \frac{1}{bm}\sum_{j=1}^{b}\sum_{k=1}^{m} X_{ijk} = \frac{1}{b}\sum_{j=1}^{b} \bar{X}_{ij\cdot}, i = 1,2,\cdots,a;$$

$$\bar{X}_{\cdot j\cdot} = \frac{1}{am}\sum_{i=1}^{a}\sum_{k=1}^{m} X_{ijk} = \frac{1}{a}\sum_{i=1}^{a} \bar{X}_{ij\cdot}, j = 1,2,\cdots,b;$$

$$n = abm$$

$$\bar{X} = \frac{1}{abm}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{m} X_{ijk} = \frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b} \bar{X}_{ij\cdot} = \frac{1}{a}\sum_{i=1}^{a} \bar{X}_{i\cdot\cdot} = \frac{1}{b}\sum_{j=1}^{b} \bar{X}_{\cdot j\cdot},$$

(平方和分解公式) 在有交互作用的双因素方差分析模型

$X_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk},\ k = 1,2,\cdots,m;\ i = 1,2,\cdots,a, j = 1,2,\cdots,b,$

中，$\mathrm{SS} = SS_A + SS_B + SS_{A\times B} + SS_e$

- $SS = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{m}\left(X_{ijk} - \bar{X}\right)^2$ **总离差平方和**，反映全体数据中的波动
- $SS_A = bm\sum_{i=1}^{a}(\bar{X}_{i\cdot\cdot} - \bar{X})^2$ **因子 *A* 的离差平方和**，反映因素A 在各个水平下的不同作用而在数据中引起的波动
- $SS_B = am\sum_{j=1}^{b}\left(\bar{X}_{\cdot j\cdot} - \bar{X}\right)^2$ **因子 *B* 的离差平方和**，反映因素B 在各个水平下的不同作用而在数据中引起的波动
- $SS_{A\times B} = m\sum_{i=1}^{a}\sum_{j=1}^{b}\left(\bar{X}_{ij\cdot} - \bar{X}_{i\cdot\cdot} - \bar{X}_{\cdot j\cdot} + \bar{X}\right)^2$ **交互效应的离差平方和**，反映由于交互效应的存在而在数据中引起的波动
- $S_e = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{m}\left(X_{ijk} - \bar{X}_{ij}\right)^2$ **误差平方和**，反映由于随机误差的作用而在数据中引起的波动

要分别检验假设

$$H_{0A}: \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0;$$
$$H_{0B}: \beta_1 = \beta_2 = \cdots = \beta_b = 0;$$
$$H_{0A\times B}: \gamma_{11} = \cdots = \gamma_{1b} = \cdots = \gamma_{a1} = \cdots = \gamma_{ab} = 0.$$

定理 在有交互作用的双因素方差分析模型 $X_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$ ，$k = 1,2,\cdots,m;\ i = 1,2,\cdots,a, j = 1,2,\cdots,b$ 中,

- $\mathrm{SS_A}$、$\mathrm{SS_B}$、$\mathrm{SS_{A\times B}}$、$\mathrm{SS_e}$ 相互独立
- $\mathrm{SS_e}/\sigma^2 \sim \chi^2(\mathrm{n-ab})$.
- 当 H_{0A} 成立, 即 $\alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$ 时, $SS_A/\sigma^2 \sim \chi^2(a-1)$;
- 当 H_{0B} 成立, 即 $\beta_1 = \beta_2 = \cdots = \beta_b = 0$ 时, $SS_B/\sigma^2 \sim \chi^2(b-1)$;
- 当 $H_{0A\times B}$ 成立, 即 $\gamma_{11} = \cdots = \gamma_{1b} = \cdots = \gamma_{a1} = \cdots = \gamma_{ab} = 0$ 时, $SS_{A\times B}/\sigma^2 \sim \chi^2((a-1)(b-1))$.

假设检验问题 $H_{0A}: \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$

取检验量为 $F_A = \frac{SS_A/(a-1)}{SS_e/(n-ab)}$,

- 当 H_{0A} 成立时, $F_A \sim F(a-1, n-ab)$;
- 当 H_{0A} 不成立时, 由于

$$E(SS_A) = bmE\left[\sum_{i=1}^{a} (\bar{X}_{i}..-\bar{X})^2\right] = bm\sum_{i=1}^{a} E[(\bar{X}_{i}..-\bar{X})^2]$$

$$= (a-1)\sigma^2 + bm\sum_{i=1}^{a} \alpha_i^2,$$

所以 SS_A 有偏大的趋势. 因此, 由 $F_A > F_\alpha(a-1, n-ab)$ 所确定的拒绝域给出了显著性水平 α 下的一个检验.

假设检验问题$\boldsymbol{H_{0B}: \beta_1 = \beta_2 = \cdots = \beta_b = 0}$

取检验量为 $F_B = \frac{SS_B/(b-1)}{SS_e/(n-ab)}$,

- 当 H_{0B} 成立时, $F_B \sim F(b-1, n-ab)$;
- 当 H_{0B} 不成立时, 由于

$$E(SS_B) = amE\left[\sum_{j=1}^{b}\left(\bar{X}_{.j.} - \bar{X}\right)^2\right] = am\sum_{j=1}^{b} E\left[\left(\bar{X}_{.j.} - \bar{X}\right)^2\right]$$

$$= (b-1)\sigma^2 + am\sum_{j=1}^{b}\beta_j^2,$$

所以 SS_B 有偏大的趋势. 因此, 由$F_B > F_\alpha(b-1, n-ab)$所确定的拒绝域给出了显著性水平 α 下的一个检验.

假设检验问题$H_{0A\times B}$: $\gamma_{11}=\cdots=\gamma_{1b}=\cdots=\gamma_{a1}=\cdots=\gamma_{ab}=0$

取检验统计量为

$$F_{A\times B}=\frac{SS_{A\times B}/(a-1)(b-1)}{SS_e/(n-ab)},$$

- 当 $\mathrm{H}_{0A\times B}$ 成立时, $F_{A\times B}\sim F((a-1)(b-1),n-ab)$;
- 当 $\mathrm{H}_{0A\times B}$ 不成立时, 由于

$$E(S_{A\times B})=(a-1)(b-1)\sigma^2+m\sum_{i=1}^{a}\sum_{j=1}^{b}\gamma_{ij}^2,$$

所以 $SS_{A\times B}$ 有偏大的趋势. 因此, 由$F_{A\times B}>F_\alpha((a-1)(b-1),n-ab)$所确定的拒绝域给出了显著性水平 α 下的一个检验.

有交互作用的双因素方差分析表

方差来源	平方和	自由度	均方和	F 值
因素A	$SS_A=bm\sum_{i=1}^{a}(\bar{X}_{i\cdot\cdot}-\bar{X})^2$	$a-1$	$MS_A=\dfrac{SS_A}{a-1}$	$F_A=\dfrac{MS_A}{MS_e}$
因素B	$SS_B=am\sum_{j=1}^{b}\left(\bar{X}_{\cdot j\cdot}-\bar{X}\right)^2$	$b-1$	$MS_B=\dfrac{SS_B}{b-1}$	$F_B=\dfrac{MS_B}{MS_e}$
交互效应$A\times B$	$SS_{A\times B}=m\sum_{i=1}^{a}\sum_{j=1}^{b}\left(\left(\bar{X}_{ij\cdot}-\bar{X}_{i\cdot\cdot}-\bar{X}_{\cdot j\cdot}+\bar{X}\right)^2\right.$	$(a-1)(b-1)$	$MS_{A\times B}=\dfrac{SS_{A\times B}}{(a-1)(b-1)}$	$F_{A\times B}=\dfrac{MS_{A\times B}}{MS_e}$
误差	$SS_e=\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{m}\left(X_{ijk}-\bar{X}_{ij\cdot}\right)^2$	$n-ab$	$MS_e=\dfrac{SS_e}{n-ab}$	
总和	$SS=\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{m}\left(X_{ijk}-\bar{X}\right)^2$	$n-1$		

例 在前例中,对于燃料与推进器的每一对水平搭配, 各发射火箭两次, 测得射程 (单位: km) 如下, 在显著性水平 $\alpha = 0.05$ 下, 分别检验各种燃料、各种推进器、交互效应是否对火箭射程有显著性影响.

燃料 \ 推进器	B_1	B_2	B_3	$\bar{x}_{i\cdot\cdot}$
A_1	158.2 152.6	152.6 141.2	165.3 160.8	155.72
A_2	149.1 142.8	154.1 150.5	151.6 148.4	149.42
A_3	160.1 158.3	170.9 173.2	139.2 140.7	157.07
A_4	175.8 171.5	158.2 151.0	148.7 141.4	157.77
$\bar{x}_{\cdot j\cdot}$	158.55	156.91	149.51	$\bar{x}=154.99$

解 按所给数据列出方差分析表($a=4,b=3,m=2,n=abm=24$) :

方差来源	平方和	自由度	均方和	F 值
因素A	261.68	3	87.23	4.42
因素B	370.98	2	185.49	9.39
交互效应$A\times B$	1768.69	6	294.78	14.93
误差	236.95	12	19.75	
总和	2638.30	23		

- 对于假设检验 H_{0A}: $\alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$, 查表得临界值 $F_{0.05}(3,12) = 3.49$<4.42, 故拒绝 H_{0A}, 即可以认为不同燃料对火箭射程有显著差异.
- 对于假设检验H_{0B}: $\beta_1 = \beta_2 = \cdots = \beta_b = 0$, 查表得临界值 $F_{0.05}(2,12) = 3.89$<9.39, 故拒绝 H_{0B}, 即可以认为不同推进器对火箭射程有显著差异.
- 对于假设检验$\mathrm{H}_{0A\times B}$: $\gamma_{11} = \cdots = \gamma_{1b} = \cdots = \gamma_{a1} = \cdots = \gamma_{ab} = 0$, 查表得临界值 $F_{0.05}(6,12) = 3.00$<14.93, 故拒绝 $\mathrm{H}_{0A\times B}$, 即可以认为交互效应显著.

由于交互效应 $A\times B$ 的 F 值为 14.93, 与因素 A、B 的 F 值相比要大得多, 所以本例的交互作用特别显著, 也就是说要注意燃料与推进器的搭配.

例 某公司欲研究3种内容的广告宣传对某种大型机械的销售量的影响，第1种广告强调运输方便性，第2种广告强调节省燃料，第3种广告强调噪音低. 在经过广告广泛宣传后，抽取了一部分订单，按订单上的订购量计算的销售量如表所示：

	第 1 季度	第 2 季度	第 3 季度	第 4 季度
广告 1	163	176	170	185
广告 2	184	198	179	190
广告 3	206	191	218	224

试根据以上资料判断广告内容不同对销售量是否有显著影响 ($\alpha = 0.05$).

填空题 2分

设置

方差来源	平方和	自由度	方差	F 值	临界值
因素	2668.17	2	1334.085		$F_{0.05}(2,9) = 4.26$
误差	1098.5	9	122.056		
总和	3766.67	11			

F= [填空1] （保留2位小数），认为广告内容不同对销售量 [填空2] (有/无)显著影响.

作答