## 模型-评价主题-统计类评价-方差分析【gyj】

**1. 模型名称**

方差分析（Analysis of Variance, ANOVA）

**2. 适用范围**

方差分析可用于在建模过程中，分析**哪些因素对模型有显著影响**，哪些没有显著影响。

换言之，就是**检验各组别间是否有差异**。

**3. 形式**

- 样本数量：两个或两个以上样本
- 数据类型：方差分析用于分析定类数据[1]与定量数据[2]之间的关系。
- 自变量个数
  - 单因素方差分析：试验中只有一个因素在改变
  - 多因素方差分析：试验中有多于一个因素在改变

**4. 求解方法**

**（1）单因素方差分析模型**

4.1.1 概念

- 方差分析的基本原理：

  从试验结果退点，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异。
  - 不同处理组的差别来源于两个：
    - 组间差异(SSb)：实验条件、不同的处理造成的差异。用变量在各组的均值与总均值之偏差平方和的总和表示。**组间自由度用dfb表示。**
    - 组内差异(SSw)：随机误差、测量误差造成的差异或个体间的差异。用变量在各组内的均值与该组内变量值之偏差平方和的总和表示。**组内自由度用dfw表示。**

4.2.1 步骤

- **根据题目写出各变量：**
  - 因素：试验中需要考察的、可以控制的条件
  - 检验指标：人们关心的实验结果
  - r：因素的水平数，即将这个因素放入几种不同的情况下检验（因素所处的状态）
  - n：样本总数，即所有水平下的子样本数之和
  - $n_i$ :一个水平下的样本数，即在这种情况下进行了几次试验。
  - $H_0$：原假设，因素对检验指标没有显著影响。
- **计算每组数据的平均值( $\bar{X}_{\cdot i}$ )**

$$\bar{X}_{\cdot i} = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}$$

- **计算全体数据的总平均值( $\bar{X}$ )**

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{n_i} X_{ij} = \frac{1}{r}\sum_{i=1}^{r}\bar{X}_{\cdot i}$$

- **计算组间平方和( $S_A$ )**

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

- **计算组内平方和( $S_E$ )**

$$S_E = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{\cdot i})^2$$

- **计算总平方和( $S_T$ )**

$$S_T = S_A + S_E$$

- **计算自由度、均方、F值、P值**:

| 方差来源 | 离差平方和 | 自由度 | 均方 | F值 | P值 |
|---|---|---|---|---|---|
| 因素A(组间) | $S_A$ | r - 1 | $\frac{S_A}{r-1}$ | $F = \frac{\frac{S_A}{r-1}}{\frac{S_E}{n-r}}$ | p |
| 误差(组内) | $S_E$ | n - r | $\frac{S_E}{n-r}$ | | |
| 总和 | $S_T$ | n - 1 | | | |

P值：F(r-1, n-r)分布**大于**F值的概率。

- **查表得到 $F_\alpha(r-1, n-1)$ 值，并与F值比较**：F值分布表
- **得出结论：**
  - 若由试验数据算得结果有 $F > F_\alpha(r-1, n-r)$ ,则拒绝原假设$H_0$
  - 或，当p < α时，拒绝原假设$H_0$
  - 【注】在方差分析中还做如下规定：
    - 如果α = 0.01时拒绝$H_0$, 则称因素A的影响**高度显著。**
    - 如果α = 0.05时拒绝$H_0$，但α = 0.01时不拒绝$H_0$，则称因素A的影响**显著。**

4.3 .1实例

题目：为了考察化工生产厂中温度对某种化工产品的收率(%)的影响，现选择了5种不同的温度。在同一温度下各做4次试验，试验结果见表7.16. 问反应温度对产品收率有无显著影响。

| | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | 55.0 | 58.0 | 57.4 | 57.1 |
| 2 | 54.4 | 56.8 | 56.0 | 56.0 |
| 3 | 54.0 | 54.1 | 54.0 | 54.0 |
| 4 | 56.4 | 57.0 | 57.0 | 57.0 |
| 5 | 56.1 | 57.0 | 54.0 | 54.0 |

- **根据题目写出各变量**：本题的因素为温度；检验指标为产品的收率；r = 5; $n_i$ = 4(表示每一个水平进行了4次测量); n = 20

  $H_0$: 温度对产品收率没有显著影响。

- **计算每组数据的平均值( $\bar{X}_{\cdot i}$ )**

$$\bar{X}_{\cdot i} = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}$$

如， $\bar{X}_{\cdot 1} = \frac{55.0+58.0+57.4+57.1}{4} = 56.875, \bar{X}_{\cdot 2} = 54.90, \bar{X}_{\cdot 3} = 54.10, \bar{X}_{\cdot 4} = 56.75, \bar{X}_{\cdot 5} = 55.80$

- **计算全体数据的总平均值( $\bar{X}$ )**

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{n_i} X_{ij} = \frac{1}{r}\sum_{i=1}^{r}\bar{X}_{\cdot i}$$

如， $\bar{X} = \frac{56.875+54.90+54.10+56.75+55.80}{5} = 55.685$ 。

- **计算组间平方和( $S_A$ )**

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

如，
$S_A = 4\cdot(56.875 - 55.685)^2 + 4\cdot(54.90 - 55.685)^2 + 4\cdot(54.10 - 55.685)^2 + 4\cdot(56.75 - 55.685)^2 + 4\cdot(55.80 - 55.685)^2 = 22.7680$

- **计算组内平方和( $S_E$ )**

$$S_E = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{\cdot i})^2$$

如，水平1中 $\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_{\cdot i})=5.1075$ ;然后按照这种方法计算出每个水平的这个值；则
$S_E=5.1075+11.32+0.06+0.27+4.86=21.6175$

- **计算总平方和( $S_T$ )**

$$S_T=S_A+S_E$$

如， $S_T=22.7680+21.6175=44.3855$

- **按公式计算自由度、均方、F值、概率**

整理得到下表：

| 方差来源 | 离差平方和 | 自由度 | 均方 | F值 | P值 |
|---|---|---|---|---|---|
| 组间 | $S_A=22.7680$ | 4 | 5.6920 | 3.9496 | 0.0220 |
| 组内 | $S_E=21.6175$ | 15 | 1.4412 | 3.9496 | 0.0220 |
| 总和 | $S_T=44.3855$ | 19 | | | |

- **查表得到 $F_\alpha(r-1,n-1)$ 值，并与F值比较**：F值分布表

$F_{0.05}(4,15)=3.60<3.9496(F)$

- **得出结论：**

由于 $F_{0.05}(4,15)=3.60<3.9496(F)$ ，拒绝$H_0$,即温度对产品收率有显著影响。

4.4 .1代码实现

Matlab(法一)

```
clc,clear,close all
a = readmatrix('data_example.txt') %注意矩阵转置
[p,t,st] = anoval(a)
Fa = finv(0.95,t{2,3},t{3,3}) %计算F分布上的α分位数
```

Matlab(法二)

```
%ANOVA方差分析
%需要输入每组每个变量的情况，会输出F值
%再将计算出的F与理论上的F分布表进行对比
%若F>F(dfb,dfw),表明各组数据存在显著性差异
%若F<F(dfb,dfw),表明各组数据不存在显著性差异
clear all;
X =[55.0000   59.0000   57.4000   57.1000
    54.4000   56.8000   52.4000   56.0000
    54.0000   54.1000   54.3000   54.0000
    56.4000   57.0000   56.6000   57.0000
    56.1000   57.0000   56.1000   54.0000];%需要处理的数据，每列为一组
[r,m]=size(X);                          %r是每组多少个变量，m是共有多少组
d1=mean(X)-mean(mean(X));               %求各组的平均值与总平均值的差距
ssb=r*sum(d1.^2);                       %求组间变异的离均差平方和
dfb=m-1;                                %计算组间变异自由度
D2=var(X,1)*r;                          %计算各组组内各个变量与组内平均值的差值的平方和的累加
ssw=sum(D2);                            %计算组内变异的离均差平方和
dfw=r*m-m;                              %计算组内变异的自由度
msb=ssb/dfb;                            %组内变异均方
msw=ssw/dfw;                            %组间变异均方
F=msb/msw                               %F值
%再将计算出的F与理论上的F分布表进行对比
%若F>F(dfb,dfw),表明各组数据存在显著性差异
%若F<F(dfb,dfw),表明各组数据不存在显著性差异
```

### (2) 多因素方差分析模型

4.2.1 概念

以双因素方差分析为例，其基本思想是：对每个因素各取几个水平，然后对各因素不同水平的每个组合做一次或若干次试验，对所的数据进行方差分析。对双因素方差分析可分为无重复试验和等重复试验两种情况。

- 无重复试验：只需检验两因素是否**分别**对指标有显著性影响。
- 等重复试验：还要进一步检验两因素是否对指标有显著的交互影响。

4.2.2 步骤

- 列双因素试验数据表：

  设A取s个水平 $A_1, A_2, \cdots, A_s$ ; B取r个水平 $B_1, B_2, \cdots, B_r$ ; 在水平组合（$B_i$, $A_j$) 下做了t 个试验。

| | $A_1$ | $A_2$ | $\cdots$ | $A_s$ |
|---|---|---|---|---|
| $B_1$ | $X_{111}, \cdots, X_{11t}$ | $X_{121}, \cdots, X_{12t}$ | $\cdots$ | $X_{1s1}, \cdots, X_{1st}$ |
| $B_2$ | $X_{211}, \cdots, X_{21t}$ | $X_{221}, \cdots, X_{22t}$ | $\cdots$ | $X_{2s1}, \cdots, X_{2st}$ |
| $\cdots$ | $\cdots$ | $\cdots$ | $\cdots$ | $\cdots$ |
| $B_r$ | $X_{r11}, \cdots, X_{r1t}$ | $X_{r21}, \cdots, X_{r2t}$ | $\cdots$ | $X_{rs1}, \cdots, X_{rst}$ |

- 写原假设：

$$H_{01}: \alpha_j = 0 (j = 1, 2, \cdots, s)$$
$$H_{02}: \beta_i = 0 (i = 1, 2, \cdots, r)$$
$$H_{03}: \gamma_{ij} = 0 (i = 1, 2, \cdots, r; j = 1, 2, \cdots, s)$$

  $\gamma$ 表示 $B_i$ 和 $A_j$ 对指标的交互影响。

- 若为**<u>无交互影响</u>**
  - 即根据经验或某种分析能够实现判定两因素之间没有交互影响，则每组试验就不必重复，t = 1.

| 方差来源 | 离差平方和 | 自由度 | 均方 | F值 |
|---|---|---|---|---|
| 因素A | $S_A$ | $s-1$ | $\frac{S_A}{s-1}$ | $F_A = \frac{S_A/(s-1)}{S_E/(s-1)(r-1)}$ |
| 因素B | $S_B$ | $r-1$ | $\frac{S_B}{r-1}$ | $F_B = \frac{S_B/(r-1)}{S_E/(s-1)(r-1)}$ |
| 误差 | $S_E$ | $(s-1)(r-1)$ | $\frac{S_E}{(s-1)(r-1)}$ | |
| 总和 | $S_T$ | $rs-1$ | | |

  - 因素A的平方和：因素A造成的组间差异( $S_A$ )

$$S_A = r \sum_{j=1}^{s} (\bar{X}_{\cdot j} - \bar{X})^2$$

  - 因素B的平方和：因素B造成的组间差异( $S_B$ )

$$S_B = s \sum_{i=1}^{r} (\bar{X}_{\cdot i} - \bar{X})^2$$

  - 随机误差（ $S_E$ ）

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

  - 总平方和（ $S_T$ ）

$$S_T = S_A + S_B + S_E = \sum_{i=1}^{r} \sum_{j=1}^{s} (X_{ij} - \bar{X})^2$$

  - 总平均值（ $\bar{X}$ ）

$$\bar{X} = \frac{1}{rs} \sum_{i=1}^{r} \sum_{j=1}^{s} X_{ij}$$

  - A组数据的平均值（ $\bar{X}_{\cdot j}$ ）

$$\bar{X}_{\cdot j} = \frac{1}{r} \sum_{i=1}^{r} X_{ij}$$

  - B组数据的平均值（ $\bar{X}_{\cdot i}$ ）

$$\bar{X}_{\cdot i} = \frac{1}{s} \sum_{j=1}^{s} X_{ij}$$

  - 判断结果：
    - $F_A < F_\alpha(s-1, (r-1)(s-1))$ 时接受 $H_{01}$ ,否则拒绝。
    - $F_B < F_\alpha(r-1, (r-1)(s-1))$ 时接受 $H_{02}$ ,否则拒绝。
- 若为**<u>交互相应</u>**:
  - 根据每个水平下做了多少次试验判断t的值。

| 方差来源 | 离差平方和 | 自由度 | 均方 | F值 |
| --- | --- | --- | --- | --- |
| 因素A | $S_A$ | $s-1$ | $\frac{S_A}{s-1}$ | $F_A=\frac{S_A/(s-1)}{S_E/[rs(t-1)]}$ |
| 因素B | $S_B$ | $r-1$ | $\frac{S_B}{r-1}$ | $F_B=\frac{S_B/(r-1)}{S_E/[rs(t-1)]}$ |
| 交互效应 | $S_{AB}$ | $(s-1)(r-1)$ | $\frac{S_{AB}}{(s-1)(r-1)}$ | $F_{AB}=\frac{S_{AB}/(r-1)(s-1)}{S_E/[rs(t-1)]}$ |
| 误差 | $S_E$ | $rs(t-1)$ | $\frac{S_E}{rs(t-1)}$ | |
| 总和 | $S_T$ | $rst-1$ | | |

- 因素A的平方和：因素A造成的组间差异( $S_A$ )

$$S_A=rt\sum_{j=1}^{s}(\bar{X_{\cdot j\cdot}}-\bar{X})^2$$

- 因素B的平方和：因素B造成的组间差异( $S_B$ )

$$S_B=st\sum_{i=1}^{r}(\bar{X_{i\cdot\cdot}}-\bar{X})^2$$

- 因素AB的平方和：因素AB造成的交互影响( $S_{AB}$ )

$$S_{AB}=t\sum_{i=1}^{r}\sum_{j-1}^{s}(\bar{X_{ij\cdot}}-\bar{X_{i\cdot\cdot}}-\bar{X_{\cdot j\cdot}}+\bar{X})^2$$

- 随机误差（ $S_E$ ）

$$S_E=\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{t}(X_{ijk}-\bar{X_{ij\cdot}})^2$$

- 总平方和（ $S_T$ ）

$$S_T=S_A+S_B+S_E+S_{AB}=\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{t}(X_{ijk}-\bar{X})^2$$

- 总平均值（ $\bar{X}$ ）

$$\bar{X}=\frac{1}{rst}\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{t}X_{ijk}$$

- AB间的交互影响（ $\bar{X_{ij\cdot}}$ ）

$$\bar{X_{ij\cdot}}=\frac{1}{t}\sum_{k=1}^{t}X_{ijk}$$

- B组数据的平均值（ $\bar{X_{i\cdot\cdot}}$ ）

$$\bar{X_{i\cdot\cdot}}=\frac{1}{st}\sum_{j=1}^{s}\sum_{k=1}^{t}X_{ijk}$$

- A组数据的平均值（ $\bar{X_{\cdot j\cdot}}$ ）

$$\bar{X_{\cdot j\cdot}}=\frac{1}{rt}\sum_{i=1}^{r}\sum_{k=1}^{t}X_{ijk}$$

- 判断结果：
  - 若 $F_A>F_{\alpha(r-1,rs(t-1))}$ ,则拒绝 $H_{01}$
  - 若 $F_B>F_{\alpha(s-1,rs(t-1))}$ ,则拒绝 $H_{02}$
  - 若 $F_{AB}>F_{\alpha((r-1)(s-1),rs(t-1))}$ ,则拒绝 $H_{03}$ ， 即认为交互作用显著

4.3.3 例子

1. 无交互影响：

题目：一种火箭使用4种燃料、3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做一次试验，得到的试验数据如下表。问各种燃料及各推进器之间有无显著差异

| | $B_1$ | $B_2$ | $B_3$ |
| --- | --- | --- | --- |
| $A_1$ | 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 |

设在显著性水平α= 0.05下检验

- 令假设 $H_1:\alpha_1=\alpha_2=\alpha_3=\alpha_4=0$；$H_2:\beta_1=\beta_2=\beta_3=0$
- 按照上述公式对每个量进行计算，汇总到表格中：

| 方差来源 | 离差平方和 | 自由度 | 均方 | F值 | P值 |
| --- | --- | --- | --- | --- | --- |
| 因素A | $S_A=157.5900$ | 3 | 52.5300 | 0.4305 | 0.7387 |
| 因素B | $S_B=223.8467$ | 2 | 111.9233 | 0.9174 | 0.4491 |
| 误差 | $S_E=731.9800$ | 6 | | | |

由于题目中规定"每个组合做一次试验"，故本题为无交互影响，t = 1

- 通过查表，得 $F_{0.05}(3,6)=4.76>F_A$ ,接受 $H_1$；$F_{0.05}(2,6)=5.14>F_B$ ,接受 $H_2$ .即，各燃料和各种推进器之间的差异对于火箭的射程没有显著影响。

4.3.4 代码：

Matlab

```
clc,clear,close all
a = readmatrix('data_1.txt')
[p,t,st] = anova2(a)
```

2. 有交互影响

题目：一种火箭使用4种燃料、3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做两次试验，得到的试验数据如下表。问各种燃料及各推进器之间有无显著差异。

| | $B_1$ | $B_2$ | $B_3$ |
| --- | --- | --- | --- |
| $A_1$ | 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 |

设在显著性水平α= 0.05下检验

- 令假设 $H_1:\alpha_1=\alpha_2=\alpha_3=\alpha_4=0$；$H_2:\beta_1=\beta_2=\beta_3=0;H_3=\gamma_{11}=\gamma_{12}=,\cdots=\gamma_{43}=0$
- 按照上述公式对每个量进行计算，汇总到表格中：

| 方差来源 | 离差平方和 | 自由度 | 均方 | F值 | P值 |
| --- | --- | --- | --- | --- | --- |
| 因素A | $S_A=261.6750$ | 3 | 87.2250 | 4.4174 | 0.0260 |
| 因素B | $S_B=370.9808$ | 2 | 185.4904 | 9.3939 | 0.0035 |
| 交互作用 | $S_{AB}=1768.6925$ | 6 | 294.7821 | 14.9288 | 0.0001 |
| 误差 | $S_E=236.9500$ | 12 | 19.7458 | | |

本题题目中规定t = 2

- 通过查表，得 $F_{0.05}(2,12)=3.89<F_A$ ,拒绝 $H_1$；$F_{0.05}(3,12)=3.49<F_B$ ,拒绝 $H_2$ . $F_\alpha(6,12)=3.00<F_{AB}$ ,拒绝$H_3$. 即，各燃料推进器之间的差异对于火箭的射程有显著影响，且交互作用显著。

4.4.2 代码

Matlab

```
clc, clear, close all
a = readmatrix('data_2.txt')
[p, t, st] = anova2(a,2)
```

**5. 补充资料**

1. 数模官网 - 方差分析
2. 数学建模算法与应用书 - p201~211
3. 全流程总结方差分析- 知乎
4. 上文中用到的其他知识点

---

1. 定类数据：数据类型为分的类别，各个类别之间无法进行比较和数学运算。例，中国的民族可分成汉族、维吾尔族、壮族等。 ↩
2. 定量数据：数字，可以进行数学运算。 ↩