《生物实验设计》
第六章　方差分析

王超

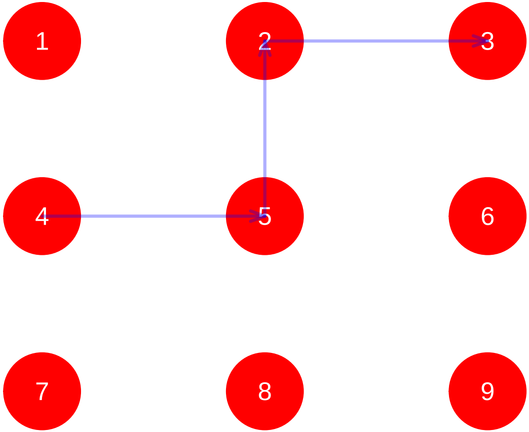
广东药科大学

Email: wangchao@gdpu.edu.cn

2022-10-02

第六章　方差分析

Check In Code: 4523



- 样本平均数的假设检验适用于样本与总体或者两个样本之间的差异显著性检验
- 实际研究中，常需要 3 个及 3 个以上样本平均数进行比较
  - 如果两两相互比较，随着样本平均数个数增加而剧增
  - n 个样本平均数需要比较的次数为 C_n^2
- 导致：
  - 检验过程繁琐
  - 无统一的试验误差，误差估计的精确性和检验的灵敏性低
  - 推断的可靠性降低

方差分析

- 方差分析 ANOVA
  - 将所有处理的观测值作为一个整体，一次比较就对所有各组间样本平均数是否有差异做出判断
  - 差异不显著，则认为他们是相同的
  - 差异显著，进一步比较是哪一组数据与其他数据不同
- 方差分析的用途
  - 多个样本平均数的比较
  - 分析多个因素间的交互作用
  - 回归方程的假设检验
  - 方差的同质性检验

- 处理效应
  - 处理因素的不同造成
- 误差效应
  - 试验过程中偶然性因素的干扰
  - 测量误差
- 方差分析的基本思想
  - 将测量数据的总变异按照变异原因不同分解为处理效应和误差效应，并作出其数量估计

- 反应测量数据变异性指标
  - 方差，即均方

$$s^2 = \frac{\sum(x-\bar{x})^2}{n-1}$$

- 分别计算出处理效应的方差和误差效应的方差，在一定显著水平下进行比较
  - 二者相差不大，说明试验处理对指标影响不大
  - 二者相差较大，说明试验处理影响是很大的，不可忽视

以单因素试验为例，假设试验考察的因素有 k 个水平，每个处理重复 n 次，共有 nk 个观测值

$$\begin{bmatrix} A_1 & A_2 & \dots & A_i & \dots & A_k \\ x_{1,1} & x_{2,1} & \dots & x_{i,1} & \dots & x_{k,1} \\ x_{1,2} & x_{2,2} & \dots & x_{i,2} & \dots & x_{k,2} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{1,j} & x_{2,j} & \dots & x_{i,j} & \dots & x_{k,j} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{1,n} & x_{2,n} & \dots & x_{i,n} & \dots & x_{k,n} \end{bmatrix}$$

处理的总和 $T_{i,:}$，

$$\begin{bmatrix} T_{1,:} & T_{2,:} & \dots & T_{i,:} & \dots & T_{k,:} \end{bmatrix}$$

平均 $\bar{x}_{i,:}$，

$$\begin{bmatrix} \bar{x}_{1,:} & \bar{x}_{2,:} & \dots & \bar{x}_{i,:} & \dots & \bar{x}_{k,:} \end{bmatrix}$$

$$x_{i,j} = \mu_i + \epsilon_{i,j}$$

- $x_{i,j}$ 表示第 i 个处理的第 j 个观测值，对于任意 $x_{i,j}$，可以用线性可加模型来进行描述
- μ_i 为第 i 个处理观测值总体平均数
- $\epsilon_{i,j}$ 为试验误差，相互独立，服从正态分布 $N(0, \sigma^2)$

单因素试验资料的数学模型

$$\mu = \frac{1}{k}\sum_{i=1}^{k}\mu_i$$

$$\tau_i = \mu_i - \mu$$

$$x_{i,j} = \mu + \tau_i + \epsilon_{i,j}$$

- μ 为总体平均数
- τ_i 为第 i 个处理的效应
- 将观测值分解为影响观测值大小的各个因素的线性组合

对 τ_i 的不同假定

- 固定模型
  - 各个处理的效应 τ_i 是固定的一个常量，由固定因素引起的效应 $\sum \tau_i = 0$
  - 除去随机误差之后每个处理所产生的效应是固定的，分析的目的在于研究 τ_i
- 随机模型
  - 各个处理的效应 τ_i 是由随机因素所引起的效应
  - τ_i 是一个随机变量，是从 $N(0, \sigma^2)$ 的正态总体中得到的
  - 研究的目的不仅是 τ_i，还有 τ_i 的变异程度
- 混合模型
  - 多因素试验中，既包括固定效应的试验因素，又包括随机效应的试验因素

- 全部观测值的变异可以用总体的方差来度量
- 方差是离均差的平方和（SS）除以自由度 $s^2 = \frac{\sum(x-\bar{x})^2}{\mathrm{d}f}$
- 依据变异来源将试验资料的总变异分解为相应的变异
- 包括总平方和（SS_T）与总自由度（df_T）的各个变异来源

(一) 平方和分解

- 引起观测值变异的原因有处理效应和误差效应
  - 处理间平均数的差异由处理效应所致
  - 同一处理内的变异由随机误差引起
- 任一观测值 x_ij 与总平均数之差可以表示为:

$$(x_{i,j} - x_{:,:}) = (x_{i,j} - \bar{x}_{i,:}) + (\bar{x}_{i,:} - x_{:,:})$$

第一节 方差分析的基本方法 三、平方和与自由度的分解

（一）平方和分解

- 两边分别平方

$$(x_{i,j} - x_{:,:})^2 = (x_{i,j} - \bar{x}_{i,:})^2 + 2(x_{i,j} - \bar{x}_{i,:})(\bar{x}_{i,:} - x_{:,:}) + (\bar{x}_{i,:} - x_{:,:})^2$$

- 每一个处理 n 个观测值离均差平方和累加，有

$$\sum_{j=1}^{n}(x_{i,j}-x_{:,:})^2 = \sum_{j=1}^{n}(x_{i,j}-\bar{x}_{i,:})^2+2\sum_{j=1}^{n}(x_{i,j}-\bar{x}_{i,:})(\bar{x}_{i,:}-x_{:,:})+\sum_{j=1}^{n}(\bar{x}_{i,:}-x_{:,:}$$

（一）平方和分解

- 因为 $\sum_{j=1}^{n}(x_{i,j}-\bar{x}_{i,:})(\bar{x}_{i,:}-x_{:,:})=(\bar{x}_{i,:}-x_{:,:})\sum_{j=1}^{n}(x_{i,j}-\bar{x}_{i,:})=0$

$$\sum_{j=1}^{n}(x_{i,j}-x_{:,:})^2=\sum_{j=1}^{n}(x_{i,j}-\bar{x}_{i,:})^2+n(\bar{x}_{i,:}-x_{:,:})^2$$

- 把 k 个处理的离均差平方再累加，得

$$\sum_{i=1}^{n}\sum_{j=1}^{n}(x_{i,j}-x_{:,:})^2=\sum_{i=1}^{n}\sum_{j=1}^{n}(x_{i,j}-\bar{x}_{i,:})^2+n\sum_{i=1}^{n}(\bar{x}_{i,:}-x_{:,:})^2$$

（一）平方和分解

总平方和 = 处理间平方和 + 处理内平方和

$SS_T = SS_t + SS_e$

$$\begin{cases} \sum_{i=1}^{n} \sum_{j=1}^{n} (x_{i,j} - \bar{x}_{i,:})^2 \\ n \sum_{i=1}^{n} (\bar{x}_{i,:} - x_{:,:})^2 \end{cases}$$

$SS_T = SS_t + SS_e$

（二）自由度分解

总自由度 = 处理间自由度 + 处理内自由度

$df_T = df_t + df_e$

$$\begin{cases} df_T = nk - 1 \\ df_t = k - 1 \\ df_e = (nk-1) - (k-1) = k(n-1) \end{cases}$$

（三）计算方差

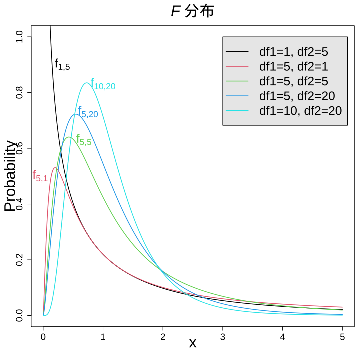
根据各变异部分得平方和与自由度，计算处理间 s_t^2 和处理内方差 s_e^2

$$\begin{cases} s_t^2 = \frac{SS_t}{df_t} \\ s_e^2 = \frac{SS_e}{df_e} \end{cases}$$

F 分布

设从一正态总体 $N(\mu,\sigma^2)$ 中随机抽样样本容量为 n_1 和 n_2 的两个独立样本，样本方差为 s_1^2 和 s_2^2，定义 F：

$$F=\frac{s_1^2}{s_2^2}$$



F 分布

- F 值的取值区间是 $[0, +\infty]$
- F 分布的平均数 $\mu_F = 1$
- F 分布曲线的性状仅取决于 df_1 和 df_2
  - 当 $df_1 = 1$ 或当 $df_1 = 2$ 时 F 分布曲线呈严重倾斜的方向 J 形
  - 当 $df_1 \geqslant 3$ 时转为左偏曲线
  - 根据 df_1 和 df_2 查 F 值分布表

- 在方差分析中进行 F 检验的目的在于推断处理间的差异是否存在
- 计算 F 时，以处理间均方 s_t^2 作分子，以处理内均方 s_e^2 作分母
- 无效假设是把各个处理的变量假设来自同一个总体，认为处理间方差与处理内方差相等

$$\begin{cases} H_0: \sigma_t^2 = \sigma_e^2 \\ H_A: \sigma_t^2 \neq \sigma_e^2 \end{cases}$$

- 无效假设是否成立，主要取决于计算出来的 F 值在 F 分布中出现的概率

方差分析不再对数据进行成对比较，而是将总体的变异进行分解，通过一次检验即完成多组处理之间的差异显著性检验

- F 检验如果否定 H_0，表明试验的变异主要来源于处理间变异，并不意味着每两个处理平均数间的差异都是显著的，也不能说明哪些平均数间有显著差异
- 要比较不同处理下平均数两两间差异的显著性，每个处理的平均数都要与其他处理的平均数进行比较
- 多重比较
    - 多个平均数两两间的相互比较

（一）最小显著差数法

- 先计算出达到差异显著的最小差数 LSD
- 然后用两个处理平均数的差值与 LSD 比较
  - $|\bar{x}_1 - \bar{x}_2| > LSD$ 在给定 α 水平上差异显著
  - $|\bar{x}_1 - \bar{x}_2| \leqslant LSD$ 差异不显著

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{\bar{x}_1 - \bar{x}_2}}$$

$$\bar{x}_1 - \bar{x}_2 = t \times s_{\bar{x}_1 - \bar{x}_2}$$

- 如果 t 值为 $t_0.05$ 或 $t_0.01$，则 $\bar{x}_1 - \bar{x}_2$ 为两个样本平均数差异达到显著或极显著水平的最小值，记为 LSD

$$\begin{cases} LSD_{0.05} = t_{0.05} \times s_{\bar{x}_1 - \bar{x}_2} \\ LSD_{0.01} = t_{0.01} \times s_{\bar{x}_1 - \bar{x}_2} \end{cases}$$

（一）最小显著差数法 (Least significant difference test)

平均数差数标准误的计算

$$s_{\bar{x}_1-\bar{x}_2}=\sqrt{\frac{s_1^2}{n_1}+\frac{s_1^2}{n_1}}=\sqrt{s_e^2(\frac{1}{n_1}+\frac{1}{n_1})}$$

如果

$$\begin{cases}|\bar{x}_1-\bar{x}_2|>LSD_{0.05}, or\\|\bar{x}_1-\bar{x}_2|>LSD_{0.01}\end{cases}$$

认为两个样本平均数的差异达显著或极显著水平

（一）最小显著差数法

- 利用 LSD 法进行多重比较，可以分 3 步
  - 计算最小显著差数 $LSD_{0.05}$ 和 $LSD_{0.01}$
  - 列出平均数的多重比较表
  - 将两两平均数的差数与 $LSD_{0.05}$ 和 $LSD_{0.01}$ 进行比较，作出统计推断

(一) 最小显著差数法

R DEMO

```
> library(agricolae)
> aloe_height <- data.frame(
+   treatment= rep(c("A1", "A2", "A3", "A4"), each = 5),
+   height = c(18.1, 18.6, 18.7, 18.9, 18.3,
+              17.4, 17.9, 17.1, 16.5, 17.5,
+              17.3, 16.9, 18.5, 18.2, 16.2,
+              15.6, 15.8, 16.7, 15.3, 16.8))
> model <- aov(height ~ treatment, data = aloe_height)
> print(LSD.test(model, "treatment", alpha = 0.01)$groups)
```

```
##    height groups
## A1  18.52      a
## A3  17.42     ab
## A2  17.28      b
## A4  16.04      c
```

（二）最短显著极差法 Duncan's new multiple range test (SSR)

R DEMO

```
> model <- aov(height ~ treatment, data = aloe_height)
> print(duncan.test(model, "treatment", alpha = 0.05)$duncan)
```

```
##      Table CriticalRange
## 2 2.997999     0.8776510
## 3 3.143802     0.9203344
## 4 3.234945     0.9470159
```

```
> print(duncan.test(model, "treatment", alpha = 0.05)$groups)
```

```
##    height groups
## A1  18.52      a
## A3  17.42      b
## A2  17.28      b
## A4  16.04      c
```

（二）q 检验 Student-Newman-Keuls test

R DEMO

```
> model <- aov(height ~ treatment, data = aloe_height)
> print(SNK.test(model, "treatment", alpha = 0.05)$snk)

##      Table CriticalRange
## 2 2.997999      0.877651
## 3 3.649139      1.068269
## 4 4.046093      1.184476

> print(SNK.test(model, "treatment", alpha = 0.05)$groups)

##    height groups
## A1  18.52      a
## A3  17.42      b
## A2  17.28      b
## A4  16.04      c
```

- 3 种检验方法检验的显著尺度关系为 $LSD \leq SSR \leq SNK$
  - 用 LSD 法检验显著的差数，用 SSR 或 SNK 法未必显著
  - 用 SNK 法检验显著的差数，用 LSD 法必然显著
- 对精度要求高的试验应用 SNK 法，一般试验用 SSR 法，试验种各个处理皆与对照相比的试验资料可用 LSD 法
- 方差分析的基本步骤：
  - 将样本数据的总平方和与总自由度分解
  - 列方差分析表进行 F 检验，分析各变异因素在总变异中的重要程度
  - 若 F 检验显著，对各处理平均数进行多重比较

- k 组资料，n 个观测值
- 方差分析表

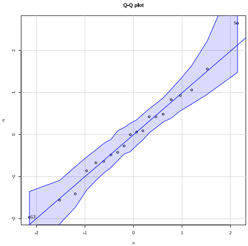
$$\begin{bmatrix} source & df & SS & s^2 & F \\ treatment & k-1 & SS_t = \frac{T_i^2}{n} - C & s_t^2 & \frac{s_t^2}{s_e^2} \\ error & k(n-1) & SS_e = SS_T - SS_t & s_e^2 & \end{bmatrix}$$

- 由于条件限制，不同处理的观测次数不同
- 方差分析表

$$\begin{bmatrix} source & df & SS & s^2 & F \\ treatment & k-1 & SS_t = \sum \frac{T_i^2}{n_i} - C & s_t^2 & \frac{s_t^2}{s_e^2} \\ error & \sum n_i - k & SS_e = SS_T - SS_t & s_e^2 & \end{bmatrix}$$

R DEMO

- 正态性检验




```
## [1]  5 13
> shapiro.test(guard_hair$length)

##
##  Shapiro-Wilk normality test
##
## data:  guard_hair$length
## W = 0.92602, p-value = 0.1294
```

R DEMO

- 单因素方差检验

```
> fit <- aov(length~region, data = guard_hair)
> summary(fit)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## region       4 173.71   43.43   50.16 1.66e-08 ***
## Residuals   15  12.99    0.87
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
```

R DEMO

- 方差齐性检验

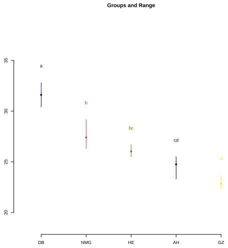
```
> bartlett.test(length~region, data = guard_hair)

##
##  Bartlett test of homogeneity of variances
##
## data:  length by region
## Bartlett's K-squared = 2.7614, df = 4, p-value = 0.5985
```

第二节　单因素方差分析

R DEMO

- **多重比较**

```
##     length groups
## DB  31.600      a
## NMG 27.400      b
## HE  26.025     bc
## AH  24.750     cd
## GZ  22.850      d
```



第三节 双因素方差分析

- 实验随机分配 60 只豚鼠，分别采用两种喂食方法（橙汁或维生素 C）
- 各喂食方法中抗坏血酸含量有三种水平（0.5mg、1mg 或 2mg）
- 每种处理方式组合都被分配 10 只豚鼠
- 观察豚鼠的牙齿长度

len	supp	dose
4.2	VC	0.5
11.5	VC	0.5
7.3	VC	0.5
5.8	VC	0.5
6.4	VC	0.5
10.0	VC	0.5

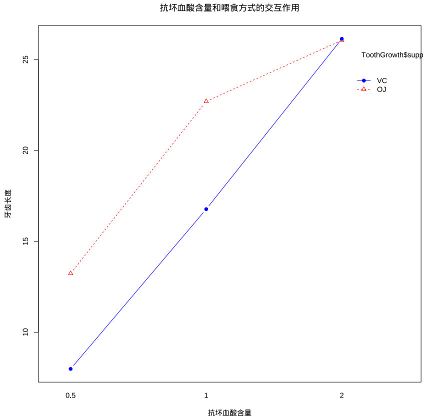
- 双因素方差分析

```
> fit <- aov(len ~ supp+dose+supp:dose, data = ToothGrowth)
> summary(fit)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## supp         1  205.4   205.4  15.572 0.000231 ***
## dose         2 2426.4  1213.2  92.000  < 2e-16 ***
## supp:dose    2  108.3    54.2   4.107 0.021860 *
## Residuals   54  712.1    13.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
```

- 交互效应的可视化

```
> interaction.plot(ToothGrowth$dose, ToothGrowth$supp,
+   ToothGrowth$len, type = "b", col = c("red", "blue"),
+   pch = c(2,19), xlab = " 抗坏血酸含量", ylab = " 牙齿长度",
+   main = " 抗坏血酸含量和喂食方式的交互作用")
```



- 数据资料体系一般都是按照试验要求事先设计好的，在试验过程中经常会因意外事件使一个或某几个数据丢失
    - 数据的缺失使平方和的线性可加模型无效
    - 无法直接进行方差分析
- 缺失的数据可用统计方法从理论上进行估计，然后进行方差分析
    - 缺失数据估计不能恢复数据，只能是补足后不致干扰其余数据
- 弥补缺失数据的原则是，使补上缺失的数据后，误差平方和最小

Table 2: one missing data

	B1	B2	B3	B4	B5	B6	B7	B8
A1	30	39	41	42	42	39	38	38
A2	37	46	x	43	51	44	35	49
A3	27	37	36	24	37	41	33	43
A4	30	42	35	40	46	47	38	46

- 根据误差平方和:

$$SS_e = SS_T - SS_A - SS_B$$

- 令 SS_e 达到最小，另 $\frac{\mathrm{d}SS_e}{\mathrm{d}x} = 0$
- 解方程得到 x

Table 3: one missing data

	B1	B2	B3	B4	B5	B6	B7	B8
A1	30	39	41	42	42	39	38	38
A2	37	46	x	43	51	44	35	49
A3	27	37	36	24	37	41	y	43
A4	30	42	35	40	46	47	38	46

- 缺失两个数据同样令 SS_e 达到最小
- 需要满足:

$$\begin{cases} \frac{\partial SS_e}{\partial x} = 0 \\ \frac{\partial SS_e}{\partial y} = 0 \end{cases}$$

- 通过求解二元一次方程求解得到 x, y

方差分析的有效性建立在一些基本假定基础上

- 正态性
  - 试验误差应当是服从正态分布 $N(0,\sigma^2)$ 的独立的随机变量
  - 每一观测值 x_{ij} 应围绕相应的平均数呈正态分布
  - 非正态分布的资料进行适当数据转换后，亦能进行方差分析
- 可加性
  - 处理效应与误差效应是可加的，并服从方差分析的数据模型
  - 试验的总变异 $x_{ij}=\mu+\alpha_i+\epsilon_{ij}$，这样才能分解为各种原因引起的变异
- 方差同质性
  - 所有试验的误差方差应具备同质性，即不同处理不能影响随机误差的方差
  - 如果发现误差异质，只要不属于研究对象本身的原因，在不影响分析正确性的条件下可将变异特别明显的数据剔除

方差同质性检验

- 假设检验都是以方差的同质性为前提，也就是各个总体的方差是相同的
- 同质性检验从各样本的方差来推断其总体方差是否相同
- 一个样本方差的同质性检验：χ^2 检验
  - $\chi^2 = \sum(\frac{x-\bar{x}}{\sigma})^2$
  - 根据 $s^2 = \frac{\sum(x-\bar{x})^2}{k-1}$，上式可以变换为 $\chi^2 = \frac{(k-1)s^2}{\sigma^2}$
  - $\chi^2 < \chi^2_\alpha$ 时，接受 $H_0: \sigma^2 = \sigma^2_\alpha$
- 一个样本方差的同质性检验：F 检验
  - 两个样本的容量、样本方差、总体方差分别为 $n_1, s_1, \sigma_1^2, n_2, s_2, \sigma_2^2$
  - 两个样本服从正态分布，且抽样是随机和独立的，其方差之比的 F 值服从 $df_1 = n_1 - 1$ 和 $df_2 = n_2 - 1$ 的 F 分布
  - $F < F_0$ 时，接受 $H_0: \sigma_1^2 = \sigma_2^2$

（一）平方根转换

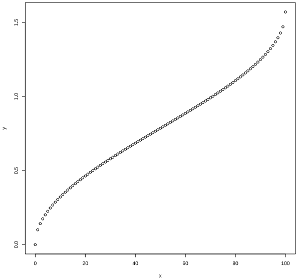
- 一些生物学观测数据为泊松而非正态分布，样本平均数和方差存在某种比例关系 $\sigma^2 = \lambda, \mu = \lambda$
- 采用平方根转换可以对方差进行降缩，减少极端大的变量对方差的影响，从而获得同质方差
- 一般直接转换成 $\sqrt{x}$，数据较小时采用 $\sqrt{x+1}$

（二）对数转换

- 如果已知资料中的效应为相乘性或非相加性，或者标准差与平均数成比例，可以使用对数转换
- 一般是将原数据转换为对数 $log(x), ln(x)$
- 使方差变成比较一致而且使效应由相乘性变成相加性
- 如果原始数据包括 0，可以采用 $log(n+1)$ 的方法
- 对数转换对于削弱大变数的作用要比平方根转换强

（三）反正弦转换

- 如果数据是比例数或以百分率表示，其分布趋向于二项分布
- 二项分布的方差与平均数之间存在一定函数关系 $\sigma^2 = npq, \mu = np$
- 转换成角度以后，接近于 0 和 1 的数值变异度增大，使方差变大，有利于满足方差同质性要求
- 方差分析时应作反正弦转换 $\theta = \sin^{-1}\sqrt{P}$
- 如果资料中的百分数为 0.3~0.7，转换对分析结果影响不大



- 对于非连续性的数据，最好在方差分析前先检查
  - 各处理平均数与相应处理内均方是否存在相关性
  - 各处理内均方间的变异是否较大
  - 如果存在以上情况，考虑对数据做转换
- 哪种方法能使处理平均数与其均方的相关性最小，哪种方法就是最合适的转换方法