

《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

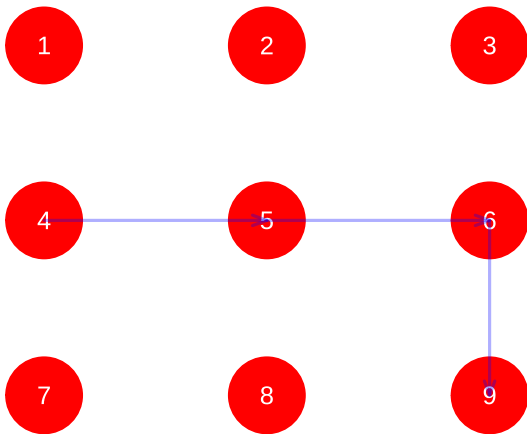
2022-09-30



廣東藥科學大學
GUANGDONG PHARMACEUTICAL UNIVERSITY

第六章 方差分析

Check In Code: 4569



- 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

第一节 方差分析的基本方法

理

一、方差分析的基本原理

- 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

第一节 方差分析的基本方法

一、方差分析的基本原理

- 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

第一节 方差分析的基本方法

二、数学模型

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 $T_{i,:}$,

$$[T_{1,:} \quad T_{2,:} \quad \dots \quad T_{i,:} \quad \dots \quad T_{k,:}]$$

平均 $\bar{x}_{i,:}$,

$$[\bar{x}_{1,:} \quad \bar{x}_{2,:} \quad \dots \quad \bar{x}_{i,:} \quad \dots \quad \bar{x}_{k,:}]$$

第一节 方差分析的基本方法

二、数学模型

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

- $x_{i,j}$ 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
- 研究的目的是不仅是 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

第一节 方差分析的基本方法

三、平方和与自由度的分解

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

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

- 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:

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

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

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

- 每一个处理 n 个观测值离均差平方和累加，有

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

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

- 因为

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

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

- 把 k 个处理的离均差平方再累加，得

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

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

总平方和 = 处理间平方和 + 处理内平方和

$$SS_T = SS_t + SS_e$$

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

$$SS_T = SS_t + SS_e$$

第一节 方差分析的基本方法

三、平方和与自由度的分解

(二) 自由度分解

总自由度 = 处理间自由度 + 处理内自由度

$$df_T = df_t + df_e$$

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

第一节 方差分析的基本方法

分解

三、平方和与自由度的

(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

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

第一节 方差分析的基本方法

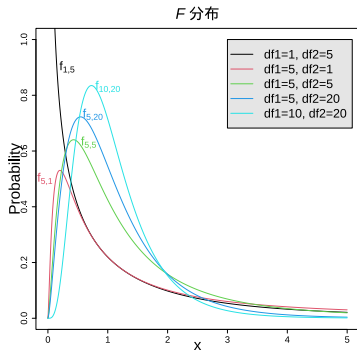
检验-- F 检验

四、统计假设的显著性

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 ：

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



第一节 方差分析的基本方法

检验-- F 检验

四、统计假设的显著性

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 $df_1 \geq 3$ 时转为左偏曲线
 - 根据 df_1 和 df_2 查 F 值分布表

第一节 方差分析的基本方法

检验-- F 检验

四、统计假设的显著性

- 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

第一节 方差分析的基本方法

五、多重比较

- 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
 - $|\bar{x}_1 - \bar{x}_2| \leq LSD$ 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

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

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

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
```

第一节 方差分析的基本方法 五、多重比较

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

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

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

第二节 单因素方差分析

一、组内观测次数相等的方差分析

- k 组资料, n 个观测值
- 方差分析表

<i>source</i>	<i>df</i>	<i>SS</i>	s^2	<i>F</i>
<i>treatment</i>	$k - 1$	$SS_t = \frac{T_i^2}{n} - C$	s_t^2	$\frac{s_t^2}{s_e^2}$
<i>error</i>	$k(n - 1)$	$SS_e = SS_T - SS_t$	s_e^2	

第二节 单因素方差分析 二、组内观测次数不相等的方差分析

- 由于条件限制，不同处理的观测次数不同
- 方差分析表

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

第二节 单因素方差分析

二、组内观测次数不相等的

方差分析

第五节 方差分析缺失数据的估计 的估计方法

一、缺失一个数据