

方差分析的假定及其检验

在方差分析模型中，假定误差项 ε 是期望值为 0、方差相等的正态独立随机变量，这些假定实际上也就是对因变量 y 的假定，在做方差分析之前，应首先对这些假定进行检验，考察数据是否适合做方差分析。本书着重介绍正态性检验和方差齐性检验。

一、正态性检验

正态性假定要求每个处理所对应的总体都服从正态分布，即对于任意一个处理，其观察值是来自正态分布总体的简单随机样本。例如，在例 7.1 中，要求每个品种的产量必须服从正态分布；在例 7.2 中，还要求每种施肥方式的产量也必须服从正态分布。

例 7.1 续/例 7.2 续 分别用 Q-Q 图和检验方法，检验各品种小麦产量是否服从正态分布 ($\alpha = 0.05$)。

解：首先，绘制 Q-Q 图来检验小麦产量是否服从正态分布。当每个处理的样本量足够大时，可以对每个样本绘制 Q-Q 图来检查每个处理对应的总体是否服从正态分布。但是，当每个处理的样本量比较小时，正态概率图中的点很少，提供的正态性信息有限时，可以将每个处理的样本数据合并绘制一幅正态概率图来检验正态性。

R 代码和结果如下文本框所示：

```

> example1<-read.csv("C:/Users/86207/Desktop/方差分析/1_1.csv")

> par(mfrow=c(1,3),cex=0.6,mai=c(0.5,0.5,0.2,0.1))

> qqnorm(example1$ 品种 1,xlab=" 期望正态值 ",ylab=" 观察值
",datax=TRUE,main="品种 1 的 Q-Q 图")

> qqline(example1$品种 1,,datax=TRUE)

> qqnorm(example1$ 品种 2,xlab=" 期望正态值 ",ylab=" 观察值
",datax=TRUE,main="品种 2 的 Q-Q 图")

> qqline(example1$品种 2,,datax=TRUE)

> qqnorm(example1$ 品种 3,xlab=" 期望正态值 ",ylab=" 观察值
",datax=TRUE,main="品种 3 的 Q-Q 图")

> qqline(example1$品种 3,datax=TRUE)##绘制例 1 中每个产品产量的正态 Q-Q 图

```

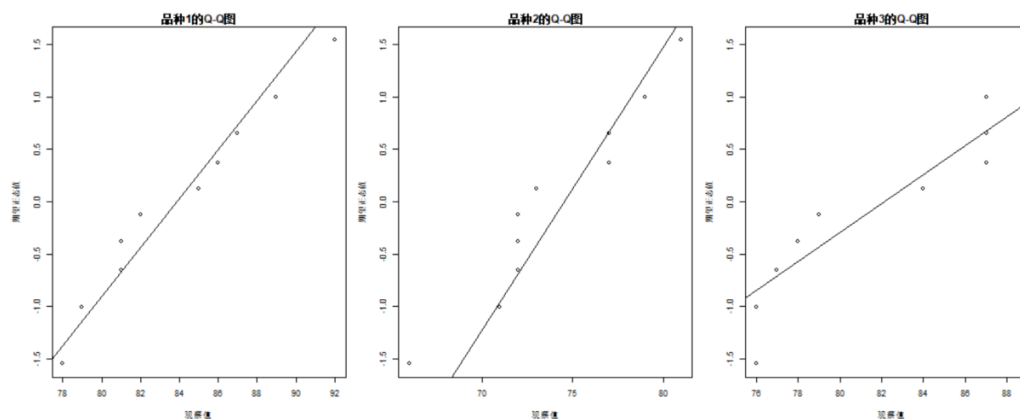


图 7-9 3 个品种产量的正态 Q-Q 图

```

> example2<-read.csv("C:/Users/86207/Desktop/方差分析/1_2.csv")
> par(cex=0.8,mai=c(0.7,0.7,0.1,0.1))
> qqnorm(example2$产量,xlab="期望正态值",ylab="观察值",data=TRUE,main="")
> qqline(example2$产量,datax=TRUE,col="red",lwd=2)
> par(fig=c(0.08,0.5,0.5,0.98),new=TRUE)
> hist(example2$产量,xlab="产量",ylab="",freq=FALSE,col="lightblue",cex.axis=
0.7,cex.lab=0.7,main="")
> lines(density(example2$产量),col="red",lwd=2)
> box()##绘制例 2 中 3 个产品产量数据合并后的正态 Q-Q 图。

```

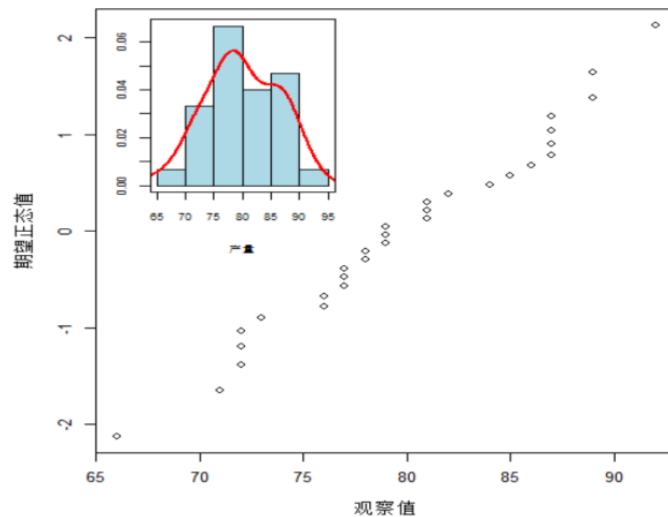


图 7-10 3 个品种产量数据合并后的正态 Q-Q 图

但由于图 7-9 对每个品种绘制的正态概率图只有 10 个数据点,很难提供正态性的证据。但从 3 个品种的产量数据合并后绘制的 Q-Q 图可以看出,产量基本上服从正态分布。

其次,分别采用 Shapiro-Wilk 和 K-S 检验来判断产量是否服从正态分布,R 代码和结果如下文本框所示:

```

> shapiro.test(example2$产量)## Shapiro-Wilk 正态性检验

Shapiro-Wilk normality test

data:  example2$产量
W = 0.97299, p-value = 0.6237

> ks.test(example2$产量,"pnorm",mean(example2$产量),sd(example2$产量))

## K-S 正态性检验

One-sample Kolmogorov-Smirnov test

data:  example2$产量
D = 0.097706, p-value = 0.9369

alternative hypothesis: two-sided

```

两种检验方法均有 $P > 0.05$ ，不拒绝 H_0 ，可以认为产量服从正态分布。需要注意的是，由于 Shapiro-Wilk 检验对正态性偏离十分敏感，当样本数据轻微偏离正态分布时，往往也会导致拒绝原假设。而方差分析对正态性的要求相对比较宽松，当正态性略微不满足时，对分析结果的影响不是很大。因此，实际中应谨慎使用这些检验。

二、方差齐性检验

方差齐性假定要求各处理的总体方差 σ^2 必须相等。比如，在例 7.1 中，要求各个品种产量的方差都相同；在例 7.2 中，要求每个品种的产量和每种施肥方式的产量方差都相同，检验方差齐性可以使用图示的方法，也可以使用标准的统计检验。

1. 图示法

检验方差齐性的图形有箱线图和残差图等。比如，绘制出每个样本数据的箱线图观察其离散程度，如果各样本箱线图的离散程度大致相同，那么方差齐性假定就可能是满足的。图 7-8 中显示了三个品种和两种施肥方式产量的箱线图，容易看出，箱线图中均没有离群点，离散程度也没有很大的差异，表明三个品种的产量和两种施肥方式的产量都可能满足齐性。

检验方差齐性的另一种图示方法是绘制残差图，残差是实际观察值与预测值的差值，残差除以残差的标准差称为标准化残差。绘制残差图的 R 代码和结果如下文本框所示：

```

>example1<-read.csv("C:/Users/86207/Desktop/方差分析/1_1.csv")
>example1=cbind(example1,id=factor(1:10))
>library(reshape)
>example1_2=melt(example1,id.vars=c("id"),variable_name="品种")
>example1_2=rename(example1_2,c(id="地块",value="产量"))
> model_1w=aov(产量~品种,data=example1_2)

> par(mfrow=c(1,2),mai=c(0.5,0.5,0.2,0.1),cex=0.6,cex.main=0.7)

> plot(model_1w,which=1)

> plot(model_1w,which=2)

```

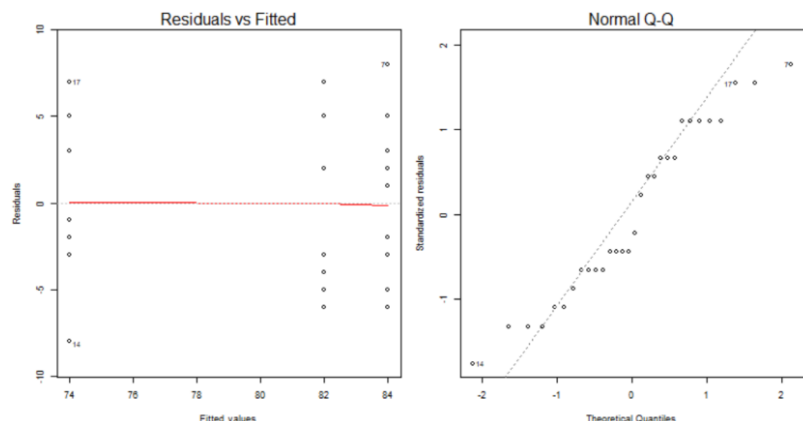


图 7-11 方差分析模型的残差图和标准化残差的 Q-Q 图

```

> example2<-read.csv("C:/Users/86207/Desktop/方差分析/1_2.csv")
> model_2wi=aov(产量~品种+施肥方式+品种:施肥方式,data=example2)
> par(mfrow=c(1,2),mai=c(0.5,0.5,0.2,0.1),cex=0.6,cex.main=0.7)
> plot(model_2wi,which=1)
> plot(model_2wi,which=2)

```

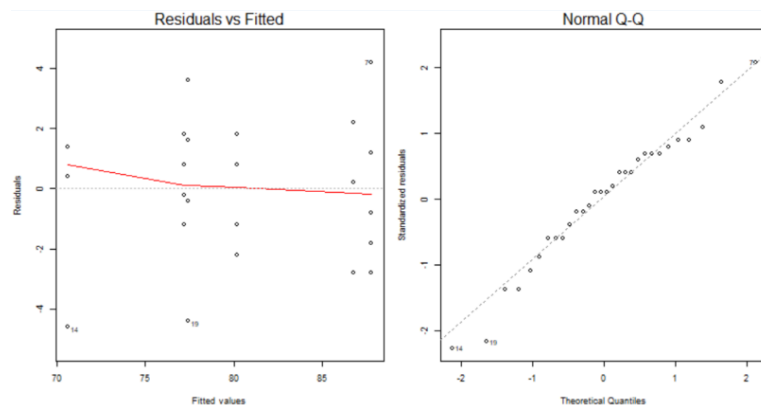


图 7-12 方差分析模型的残差图和标准化残差的 Q-Q 图

图 7-11 左图是残差图，图中的横坐标是模型的拟合值，纵坐标是预测残差。可以看出，3 个品种的残差都没有离群点，拟合值和残差的散点随机分布在一个水平带之内，而且其离散程度也基本上一样，这表示满足方差齐性的假定，此外，该图也可以用于评价方差分析模型的拟合效果，如果模型拟合得很好，那么拟合值与残差值的观测点就应该在一条水平线附近波动。从图 7-11 左图可以看出，产量与品种之间的线性关系假定成立，表示所选的方差分析模型是合理的。

图 7-11 右图是标准化残差的正态 Q-Q 图，该图的作用与图 7-10 类似，可用于检验残差的正态性假定是否成立。从该图可以看出，各个点基本上在直线周围分布，没有固定模式，可以认为 ε 的正态性假定基本成立，图 7-12 的解读类似，容易看出，不同品种和不同施肥方式的产量基本上都满足方差齐性。

2. 检验法

当各处理的样本量较小时，利用图示方法很难发现离散程度的差异，这时可以使用检验方法。

对于自变量的 I 个处理，方差齐性检验的假设是：

$$\begin{aligned} H_0 : \sigma_1^2 &= \sigma_2^2 = \dots = \sigma_I^2 \\ H_1 : \sigma_1^2, \sigma_2^2, \dots, \sigma_I^2 &\text{至少两个不同} \end{aligned}$$

(1) Bartlett 方差齐性检验。

如果所有 I 个总体都近似服从正态分布，可以使用 Bartlett 检验。当各样本量相等时，该检验的统计量为：

$$B = \frac{(n-1) \left[I \ln \bar{s}^2 - \sum s_i^2 \right]}{1 + \frac{I+1}{3I(n-1)}} \quad (7-15)$$

式中， $n = n_1 = \dots = n_I$ ； s_i^2 是第 i 个样本的方差； $\bar{s}^2 = \frac{\sum s_i^2}{I}$ ，即 I 个样本方差的平均， $\ln x$ 为 x 的自然对数。

当各样本量不相等时，该检验统计量为：

$$B = \frac{\left[\sum (n_i - 1) \right] \ln \bar{s}^2 - \sum (n_i - 1) \ln s_i^2}{1 + \frac{1}{3(n-1)} \left\{ \sum \frac{1}{(n_i - 1)} - \sum \frac{1}{(n_i - 1)} \right\}} \quad (7-16)$$

式中, n_i 为第 i 各样本的样本量; s_i^2 为第 i 个样本的方差; $\bar{s}^2 = \frac{\sum (n_i - 1)s_i^2}{\sum (n_i - 1)}$, 即 I

个样本方差的加权平均; $\ln x$ 为 x 的自然对数。

当原假设成立时, 上述统计量服从自由度为 $(I-1)$ 的 χ^2 分布。若 $B > \chi_{\alpha}^2$ 或 $P < \alpha$, 拒绝 H_0 , 表示各总体的方差不相等。

例 7.2 续 (2) 用 Barlett 方法检验不同品种的产量和不同施肥方式的产量是否满足方差齐性($\alpha = 0.05$)。

解: Barlett 方差齐性检验的 R 代码和结果如下文本框所示:

```
> bartlett.test(产量~品种,data=example2)

Bartlett test of homogeneity of variances
data: 产量 by 品种
Bartlett's K-squared = 0.30152, df = 2, p-value = 0.8601

> bartlett.test(产量~施肥方式,data=example2)

Bartlett test of homogeneity of variances
data: 产量 by 施肥方式
Bartlett's K-squared = 0.42431, df = 1, p-value = 0.5148
```

由于 P 值均大于 0.05, 不拒绝 H_0 , 可以认为不同品种的产量和不同施肥方式的产量均满足方差齐性。

(2) Levene 方差齐性检验

如果所有 I 个总体均不服从正态分布, 可以使用 Levene 检验, 它是由 H.Levene 于 1960 年提出的。该检验既可以用于正态总体, 也可以用于非正态总体。检验统计量为:

$$F = \frac{MSA}{MSE} \sim F(I-1, n-1) \quad (7-17)$$

式中, MSA 和 MSE 是对因变量实施 $y'_i = |y_i - Med_i|$ 的变换后进行方差分析得到的处理均方和残差均方; Med_i 为第 i 个处理的 y 的中位数。

如果 $P < \alpha$, 拒绝 H_0 , 表示各总体的方差不相等。

例 7.2 续 (3) 用 Levene 方法检验不同品种的产量和不同施肥方式的产量是否满足方

差齐性($\alpha = 0.05$)。

解：以 3 个品种的产量为例说明计算过程，3 个品种产量的中位数分别为： $Med_1=83.5$ ， $Med_2=72.5$ ， $Med_3=81.5$ 。对每个品种的产量分别减去其中位数后取绝对值，再进行方差分析，得到 $MSA = 5.2$ 。由式(7-17)得到的统计量为：

$$F = \frac{5.2}{5.596296} = 0.929186$$

相应的 $P = 0.407138$ ，不拒绝 H_0 ，可以认为小麦产量满足方差齐性。

Levene 检验的 R 代码和结果如下文本框所示：

```
> library(car)
> leveneTest(产量~品种,data=example2)

Levene's Test for Homogeneity of Variance (center = median)
      Df  F value  Pr(>F)
group  2   0.9292  0.4071
      27

> leveneTest(产量~施肥方式,data=example2)

Levene's Test for Homogeneity of Variance (center = median)
      Df  F value  Pr(>F)
group  1   0.2323  0.6336
      28
```

由于 P 值均大于 0.05，不拒绝 H_0 ，可以认为各个品种的产量和不同施肥方式的产量满足方差齐性。

需要提醒的是，在方差分析中，对方差齐性的要求相对比较宽松，当方差略有不同时，对分析结果的影响不是很大，特别是当各处理的样本量相同时，方差分析对不等方差是稳健的。

除正态性和方差齐性假定外，方差分析中还有一个重要的假定，即独立性。该假定要求每个样本数据是来自不同处理的独立样本，比如在例 7.2 中，三个品种的产量数据是来自三个不同品种产量总体的三个独立样本，两种施肥方式的产量是来自两种施肥方式产量总体的两个独立样本。方差分析对独立性的要求比较严格，若该假设得不到满足，方差分析的结果往往

会受到较大影响。独立性可在实验设计之前予以确定，不需要检验，因为在获取数据之前，实验的安排是否独立，研究者本身就是清楚的。