方差分析

王 成[1]
上海交通大学数学系

June 9, 2015

[1]chengwang@sjtu.edu.cn

ANOVA

方差分析（Analysis of Variance, ANOVA): 主要用来检测不同组之间是否有差异， 是R.A.Fister首先提出。是统计学里一个常用的方法，一般用于两个及两个以上样本均数差别的显著性检验。

引例

例如,我们对全国几个主要城市的收入感兴趣，经过抽样之后会有这样的数据：

$$
\begin{array}{lllll}
\text{北京:} & X_{11} & X_{12} & \cdots & X_{1n_1} \\
\text{上海:} & X_{21} & X_{22} & \cdots & X_{2n_2} \\
\text{广州:} & X_{31} & X_{32} & \cdots & X_{3n_3} \\
\text{深圳:} & X_{41} & X_{42} & \cdots & X_{4n_4} \\
 & & \vdots & &
\end{array}
$$

注意数据量$n_1, n_2, \ldots$可能不相等。 我们感兴趣的问题是：
这几个城市的人均收入是否有显著差异？如果是，那些城市与其他显著不同？

例 6.1.1

某食品公司对一种食品设计了四种不同的包装，为了考察哪种包装最受顾客欢迎，选取了四个商店进行跟踪测试，数据如下：

包装A1:	12	18	11	14
包装A2:	14	12	13	12
包装A3:	19	17	21	18
包装A4:	24	30	23	21

类似的，这里也可以有更多地商店数据。下面代码展示了四个包装对应的销售均值：

```
## [1] 13.75 12.75 18.75 24.50
```

方差分析模型

刚刚的两个例子涉及到的都是单因素方差分析，一般的我们可以假设某个因素有g个分组，对应的样本数据为：

Group 1:	X_{11}	...	X_{1n_1}
Group 2:	X_{21}	...	X_{2n_2}
...	...	...	...
Group g:	X_{g1}	...	X_{gn_g}

对于每个组，我们可以假定数据满足正态分布，既

X_{11}	...	X_{1n_1}	$\sim N(\mu_1, \sigma^2)$
X_{21}	...	X_{2n_2}	$\sim N(\mu_2, \sigma^2)$
...	...	...	
X_{g1}	...	X_{1n_g}	$\sim N(\mu_g, \sigma^2)$

或者我们可以用类似回归分析模型的表示方式：

$$X_{ij} = \mu_i + \epsilon_{ij},\ i = 1, \ldots, g,\ j = 1, \ldots, n_i,$$

这里的$\epsilon_{ij} \sim N(0, \sigma^2)$. 方差分析关心的问题是：

$$H_0 :\ \mu_1 = \cdots = \mu_g. \tag{1}$$

这里g组可以理解为因素的g个水平，而μ_i 为因素不同水平下的实验指标，所以方差分析其实研究的就是不同水平下是否有差异化的假设检验问题。

方差分析与假设检验

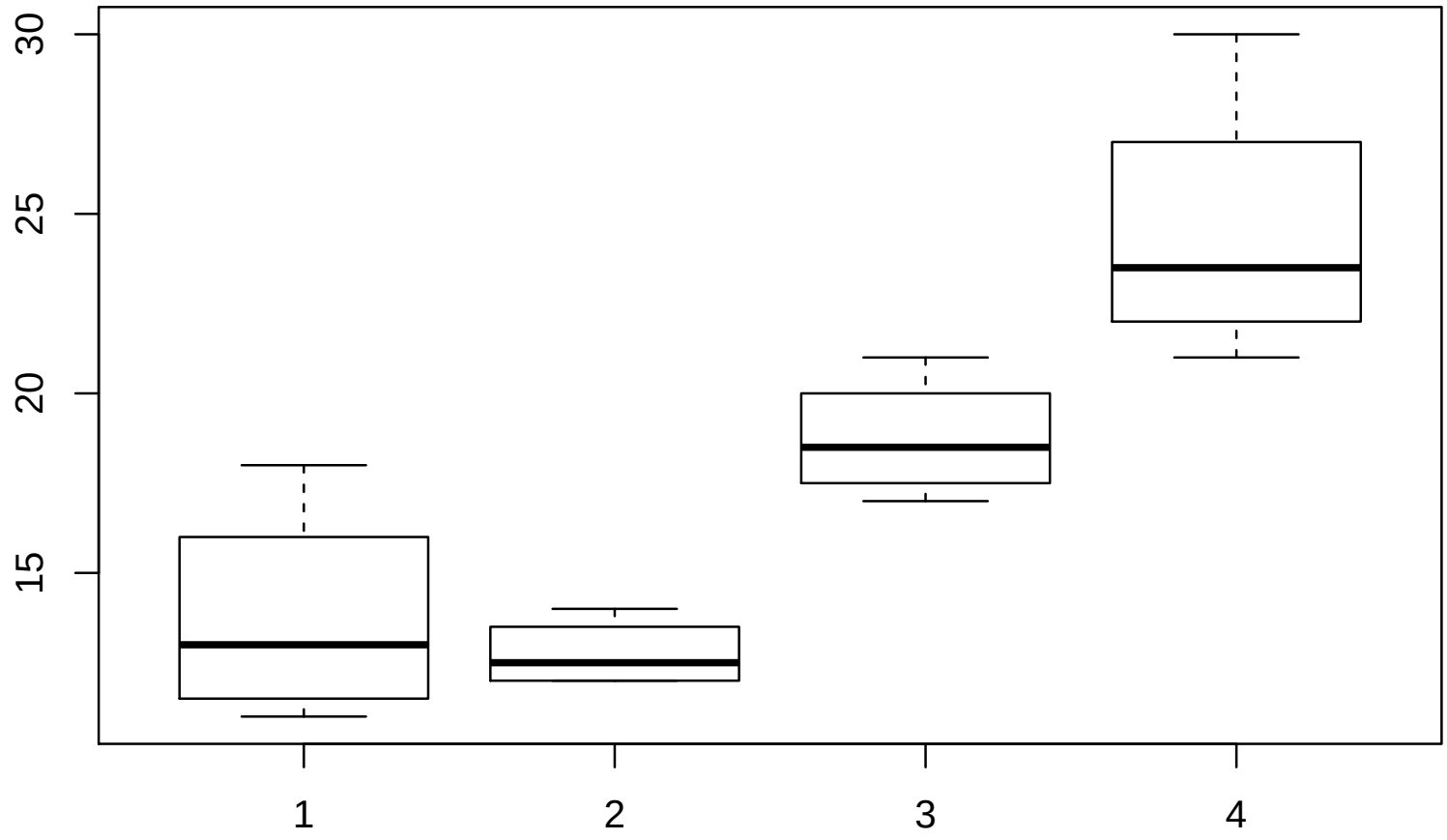
如果特别的$g = 2$, 方差分析问题就是$H_0 : \mu_1 = \mu_2$, 这也就是两样本均值检验问题。 而对于一般的$g > 2$, 方差分析问题也可以理解成很多个（小的）假设检验问题：

$$Q_{ij} \quad H_0 : \mu_i = \mu_j, \ i, j = 1, \ldots, g. \tag{2}$$

当然，我们可以检验每一个Q_{ij}。如果所有的都成立，那么对应的认为原始的多个假设检验成立。 对于例 6.1.1， 我们有：

```
t.test(x[1,],x[2,])


##
##  Welch Two Sample t-test
##
## data:  x[1, ] and x[2, ]
## t = 0.61721, df = 3.5687, p-value = 0.5743
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.721086  5.721086
## sample estimates:
## mean of x mean of y
##     13.75     12.75
```


30
25
20
15
1
2
3
4

把四个包装对应的销售数据做两两地均值检验，我们有对应的p-value结果如下：

$(1,2):\ 0.5743$　　$(1,3):\ 0.03968$　　$(1,4):\ 0.005465$

$(2,3):\ 0.00206$　　$(2,4):\ 0.007022$　　$(3,4):\ 0.05144$

所以在95% 水平下，不拒绝 $\mu_1 = \mu_2$，$\mu_3 = \mu_4$。在99% 水平下，不拒绝 $\mu_1 = \mu_2, \mu_1 = \mu_3, \mu_3 = \mu_4$.

注： 把一个假设检验分拆成多个，会造成误差积累，并不能真正替代原始的问题。

方差分析原理 1

对于每一组数据 $X_{i1}, \ldots, X_{in_i}$，我们考虑组内的样本均值和样本方差：

$$\bar{X}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij},\ S_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2,$$

以及全体数据的样本均值和样本方差：

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

对应组内样本均值和总体样本均值，我们可以把观察样本X_{ij} 分解成

$$X_{ij} = \bar{X} + (\bar{X}_i - \bar{X}) + (X_{ij} - \bar{X}_i),$$

即 观测值 $=$ 总的样本平均 $+$ 组间样本均值差 $+$ 残差.

方差分析原理 2

对应的，我们可以考虑以下几个统计量：

$$\text{总平方和:} S_T^2 = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(X_{ij} - \bar{X})^2,$$

$$\text{组间差平方和:} S_A^2 = \sum_{i=1}^{g} n_i(\bar{X}_i - \bar{X})^2,$$

$$\text{组内差平方和:} S_E^2 = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2$$

其中，我们有 **平方和分解公式**

$$S_T^2 = S_A^2 + S_E^2. \tag{3}$$

构造统计量

对于组内差，我们有 $\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_i)^2/\sigma^2 \sim \chi^2_{n_i-1}$，再由每组数据相互独立，对于组内差平方和，我们有

$$S_E^2/\sigma^2 \sim \chi^2_{n-g}. \tag{4}$$

而对于总平方和，当原假设成立时候，即$\mu_1 = \cdots = \mu_g$，所有观察数据可以视为$N(\mu_1, \sigma^2)$的观察样本，所以，$S_T^2/\sigma^2 \sim \chi^2_{n-1}$.
对于组间差平方和，我们有 $S_A^2/\sigma^2 \sim \chi^2_{g-1}$。

单因素方差分析检验统计量

当原假设不成立时候，S_A^2 应该很大。反过来，原假设成立时候S_A^2应该很小，所以检验统计量可以基于S_A^2。 为了消除未知的σ^2, 我们考虑F统计量：

$$F = \frac{S_A^2/(g-1)}{S_E^2/(n-g)}. \tag{5}$$

在原假设成立时候，我们有 $F \sim F_{g-1,n-g}$. 在给定显著水平α下，拒绝域为

$$K_0 = \{F > F_{1-\alpha}(g-1, n-g)\}. \tag{6}$$

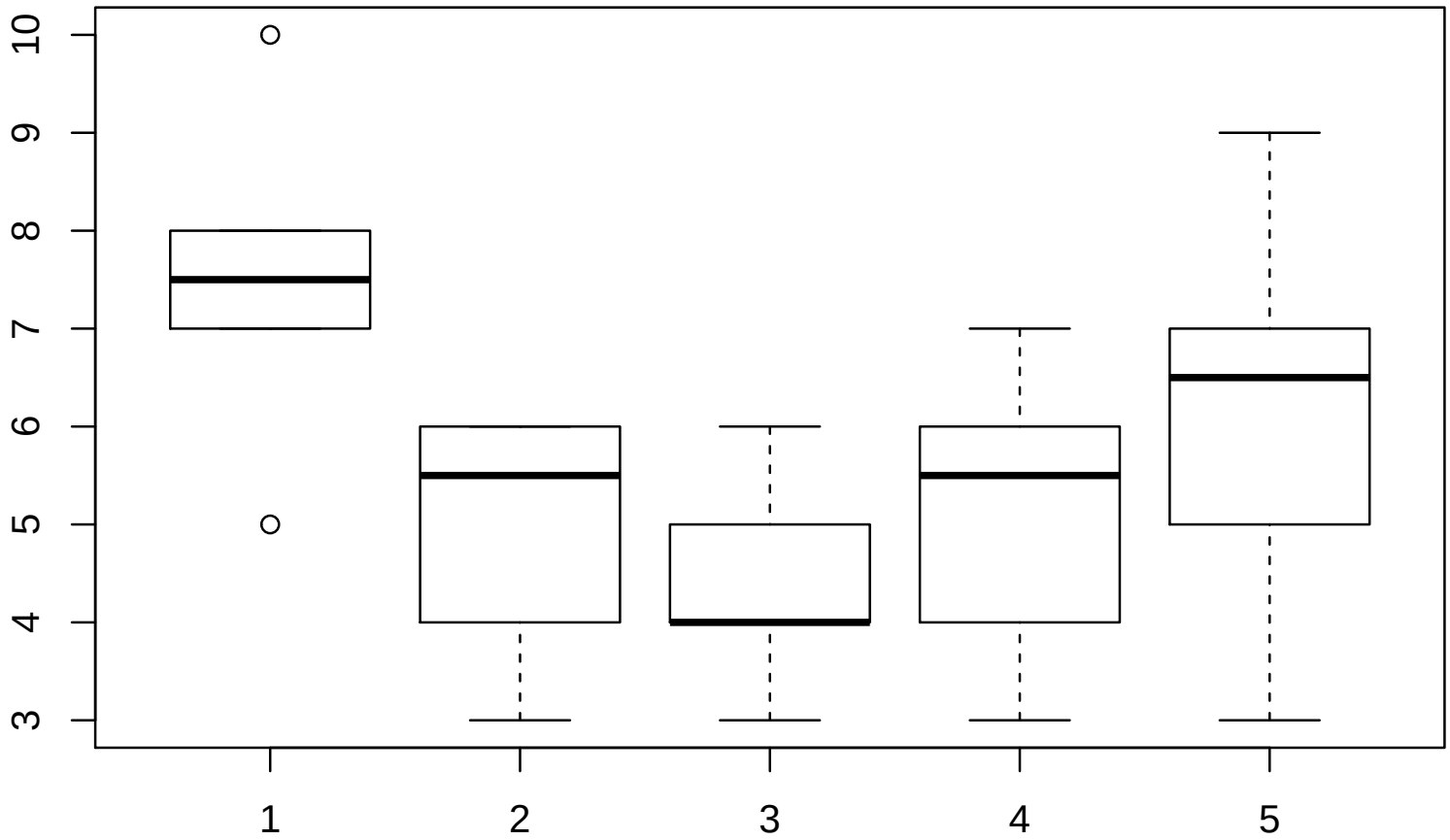
例 6.1.1 方差分析结果

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## group      3 348.69 116.229  16.361 0.0001537 ***
## Residuals 12  85.25   7.104
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

例 6.2.1

设有五种治疗某疾病的药物，为了对比去功效差异，追踪了30名病人，将他们分成5组，记录了病人从开始服药到痊愈所需要的时间 ...

例 6.2.1



例 6.2.1 方差分析结果

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value  Pr(>F)
## group      4 36.467  9.1167   3.896 0.01359 *
## Residuals 25 58.500  2.3400
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

多因素方差分析

我们首先以两因素为例：

$$(i,j)\ X_{ij1},\ldots,X_{ijn_{ij}},$$

这里 $i=1,\ldots,r$, $r=1,\ldots,s$.
对于每个组，我们可以假定数据满足正态分布，既

$$X_{ij1},\ldots,X_{ijn_{ij}} \sim N(\mu_{ij},\sigma^2)$$

回归数据的表示方式：

$$X_{ijk} = \mu_{ij} + \epsilon_{ijk},\ i = 1, \ldots, r,\ j = 1, \ldots, s, k = 1, \ldots, n_{ij}$$

这里的$\epsilon_{ijk} \sim N(0, \sigma^2)$. 我们关心的问题是：

$$H_0 :\ \mu_{11} = \cdots = \mu_{rs}. \tag{7}$$

类似的可以构造出组间差平方和和组内差平方和以及对应的F统计量等，感兴趣的读者可以自己查阅相关资料。

Thank you!