简单在 R 中做方差分析

姜晓东

2015

在正态分布的假设下，比较两组计量资料数据采用 t 检验；同时比较多组数据则要采用方差分析。所有的统计软件，如SPSS、Stata、SAS，或带有部分统计功能的软件，如Origin、Excel等都可以做单因素方差分析。R 语言是专业的统计工具，并可以免费下载安装并易于使用。本文介绍在 R 中进行方差分析。

有如下数据[1] 在 Excel 中，每列为 1 组，共 4 组。我们要比较各组差异。

A	B	C	D
1600	1500	1640	1510
1610	1640	1550	1520
1650	1400	1600	1530
1680	1700	1620	1570
1700	1750	1640	1640
1700		1600	1600
1780		1740	
		1800	

首先，我们要把数据另存为 csv 格式，本例中我们存为 “1.csv”。这样就方便在 R 中打开。

其次，我们启动 R 程序，在**文件**菜单中设置 “工作目录” 为数据文件所在的目录。然后在命令行中输入命令读入文件：

```
mydat=read.csv("1.csv", header=FALSE); # header 参数决定是否把文件第一行作为标题名。
```

这样，数据就保存在变量 mydat 中了。另外，设置 header 参数是为了防止第一行数据变成标题名。我们可以在命令窗口中输入 mydat 来查看其中的内容。你可以在命令窗口看到，数据中有一些 NA 值（Not a Number, NA），这些是由于我们的原始数据每组长度不一样，有空白所致。

接下来，我们要做的就是重新整理数据的格式，把所有的数值放在一列，把分组信息放在另外一列；简单地，我们使用 R 中的 reshape2 包来做这件事情。

```
library(reshape2); # 载入包，如果以前没装过，需要 install.packages("reshape2") 命令安装。
mynew=melt(mydat); # 重新整理数据，把数值放在一列 (value)，分组信息在另外一列（variable)。
```

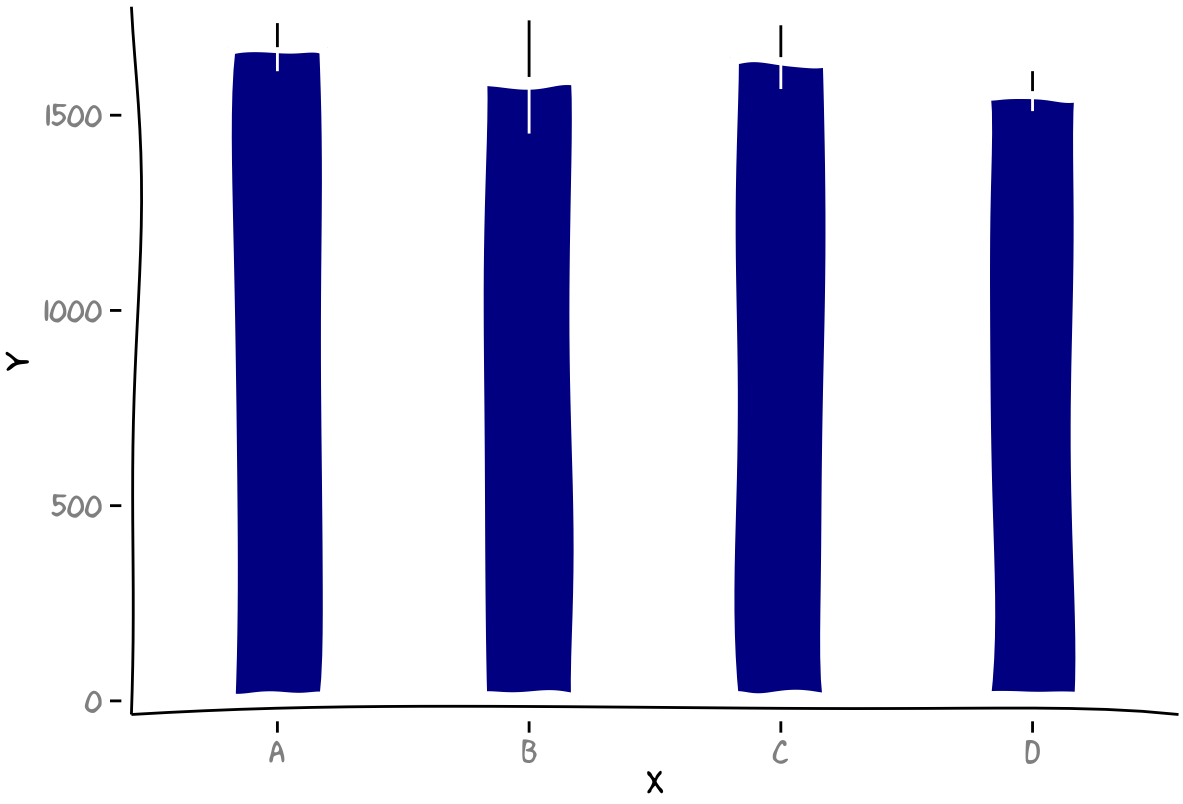
通过查看，可以知道，数值和分组信息分别放在变量 mynew 的 value 列和 variable 列中了。接下来，就可以进行方差分析了：

```
ret=aov(value~variable, data=mynew); # 方差分析，分组信息是自变量，数值是因变量
summary(ret); # 汇总输出结果
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## variable     3  49212   16404   2.166  0.121
## Residuals   22 166622    7574
## 6 observations deleted due to missingness
```

可以看到 p 值（Pr）很大，方差分析没有统计学意义。显示数据不存在组间差异。

其实，通过简单地画一下数据图，也可以看出来大致的趋势。我们有必要通过图的直观性反过来验证一下统计计算的可靠性，避免出现一些低级错误。



当方差分析有意义时，可以进一步在各组间多重比较：

```r
pairwise.t.test(mynew$value, mynew$variable)  # 各组多重比较
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  mynew$value and mynew$variable
##
##    V1   V2   V3
## V2 0.59 -    -
## V3 1.00 0.95 -
## V4 0.18 1.00 0.39
##
## P value adjustment method: holm
```

可以看到各组比较的结果，pairwise 默认采用 Holm 方法，这个是改进的 Bonferroni 法。比较合适于多数情况。

当各组数据数目大致相等时，也可以选用国内用的比较多的 Tukey 法，可以在命令行中输入 TukeyHSD(ret) 来计算 p 值，异其中 ret 是方差分析的返回值，计算的结果与以上方法比较是相近的。

R 语言计算方差分析，看似繁琐，但区区只有几行语句，就可实现。并且可以写成脚本，便于批量处理数据。R 语言中，稍微改变参数就可以进行多因素方差分析等计算，非常方便。

关于 R 的入门语法，有一份翻译成中文的官方文档可以下载[2]。关于在 R 中进行各种统计分析，可以参考薛毅编写的《统计建模与 R 软件》[1]，此书非常全面。

在学习 R 语言的过程中，如有疑问可以在**统计之都** (cos.name)上搜索或发帖求助。这个网站的常驻人群主要是统计学专业的研究生或博士。(该网站在我们这里有时会打不开，需要反复刷新，但在电信网却正常。)

参考文献

1. 薛毅，陈立萍，《统计建模与 R 软件》，清华大学出版社。
2. R Development Core Team, http://cran.r-project.org/doc/contrib/Ding-R-intro_cn.pdf