Data Analysis Using R: Chapter09

罗智超 (ROKIA.ORG)

1 通过本章你将学会

- 描述统计分析
- 列联分析
- 相关分析
- 分布拟合

2 描述统计分析

```
vars<-c("mpg","hp","wt")
summary(mtcars[vars])
boxplot(scale(mtcars[vars]))
hist(mtcars[,1])
# 自定义描述统计函数
mystat<-function(x,na.omit=FALSE){
  if (na.omit) x<-x[!is.na(x)]
  m<-mean(x)
  n<-length(x)
  s<-sd(x)
  skew<-sum((x-m)^3/s^3)/n
  kurt<-sum((x-m)^4/s^4)/n-3
  return(c(n=n,mean=m,stdev=s,skew=skew,kurtosis=kurt))

}
```

```
sapply(mtcars[vars],mystat)
mystat(mtcars$mpg)
```

- 分组计算

```
# 方法一 aggregate()
aggregate(mtcars[vars],by=list(am=mtcars$am),mean)
#aggregate() 只能使用 mean 这样返回一个值的函数

# 方法二 by(), 返回 list
vars<-c("mpg","hp","wt")
#dstats<-function(x)(c(mean=mean(x),sd=sd(x)))
dstats <- function(x)
  apply(x,MARGIN = 2,function(x)
    return(c(mean = mean(x),sd = sd(x))))
by(mtcars[vars], mtcars$am, dstats)

# 方法三 doBy::summaryBy(),psych::describe.by()
# 方法四 reshape

library(reshape)
dstats<-function(x)(c(n=length(x),mean=mean(x),sd=sd(x)))
dfm<-melt(mtcars,measure.vars=c("mpg","hp","wt"),id.vars=c("am","cyl"))
cast(dfm, am + cyl + variable ~ ., dstats)

# 方法四：dplyr::group_by
```

- 频数表和列联表

```
library(vcd)
head(Arthritis)
# 生成列联表常用函数
table(var1,var2, ) # 使用 n 个类别型变量（因子）创建一个 n 维列联表
```

```
xtab(formula,data)# 使用一个公式和一个矩阵或数据框创建一个 n 维列联表
prop.table(table,margins)# 依 margins 定义的边际列表将表中条目表示为分数形式
margin.table(table,margins)# 依 margins 定义的边际列表计算表中条目的和
addmargins(table,margins)# 将概述边 margins(默认是求和结果)放入表中
ftable(table)# 创建一个紧凑的“平铺”式列联表
```

– 一维列联表

```
mytable<-with(Arthritis,table(Improved))
prop.table(mytable)
```

– 二维列联表

```
# 方法一
mytable<-table(a,b)
# 方法二
mytalbe<-xtable(~a+b,data=mydata)
# 方法三 gmodels::CrossTable()

# 举例
mytable<-xtabs(~Treatment+Improved,data=Arthritis)
margin.table(mytable,1)
prop.table(mytable,1)
addmargins(mytable,FUN = sum)
```

– 多维列联表

```
mytable<-xtabs(~Treatment+Sex+Improved,data=Arthritis)
ftable(mytable)
#mosaicplot()
```

3 独立性检验

- 卡方独立性检验

卡方检验是一种用途很广的计数资料的假设检验方法。它属于非参数检验的范畴，主要是比较两个及两个以上样本率 (构成比) 以及两个分类变量的关联性分析。其根本思想就是在于比较理论频数和实际频数的吻合程度或拟合优度问题。

它在分类资料统计推断中的应用，包括：两个率或两个构成比比较的卡方检验；多个率或多个构成比比较的卡方检验以及分类资料的相关分析等。

```
library(vcd)
mytable<-xtabs(~Treatment+Improved,data=Arthritis)
chisq.test(mytable)
addmargins(mytalbe)
```

- Fisher 精确检验

Fisher 精确性检验的使用条件是样本量 n<40 或者理论频数 T<1

- Cochran-Mantel-Haenszel 检验

我们通常成 CMH 检验，是临床试验中分类型数据最常用的方法之一，它可以对一些分层变量进行调整，从而获得反应率的总体比较. 最为最为常见的应用是在多中心试验中对研究中心进行调整而进行两组率的比较.

原假设是两个名义变量在第三个变量的每一层中都是条件独立的。

```
mytable<-xtabs(~Treatment+Improved+Sex,data=Arthritis)
mantelhaen.test(mytable)
```

- 相关性度量

拒绝了变量间相互独立的原假设，就要衡量相关性强弱。vcd 包中的 assocstats() 函数计算二维列联表的 phi 系数，列联表系数和 Cramer's V 系数。较大的系数值意味着较强的相关性。vcd 包中的 kappa() 函数可以计算混淆矩阵 (confution matrix) 的 Cohen's Kappa 值以及加权的 kappa 值，kappa 系数是一种计算分类精度的方法。混淆矩阵可以表示两位评委对同一系列对象进行分类所得结果的一致程度。

```
library(vcd)
mytable<-xtabs(~Treatment+Improved,data=Arthritis)
assocstats(mytable)
```

4 相关

相关系数用来描述定量变量之间的关系。

4.1 相关类型

4.1.1 Pearson,Spearman 和 Kendall 相关

Pearson 衡量两个定量变量之间的线性相关程度 Spearman 衡量分级定序变量之间的相关程度 Kendall's Tau 相关系数是一种非参数的等级相关度量

```
#state.x77
cor(x,use=,method=)
#user=all.obs,everything,complete.obs,pairwise.complete.obs
#method=pearson,spearman,kendall

cov()
```

4.1.2 偏相关

在控制一个或者多个变量时，另外两个定量变量之间的相互关系。

```
library(ggm)
pcor(x1,x2)
```

4.2 相关性检验

```
#Method1
cor.test(x,y,alternative = ,method=)
#Method2
```

```
library(psych)
corr.test()
```

5 t 检验

5.1 独立样本的 t 检验

```
#MSASS:UScrime
library(MASS)
t.test(Prob~So,UScrime)
```

5.2 非独立样本的 t 检验

```
library(MASS)
sapply(UScrime[c("U1","U2")],function(x)(c(mean=mean(x),sd=sd(x))))
with(UScrime,t.test(U1,U2,paired = T))
```

多于两组之间的比较可以使用方差分析

6 组间差异的非参数检验

```
# 如果两组数据独立，可以使用 Wilcoxon 秩和检验（Mann-Whitney
#U 检验），来评估观测是否是从相同的概率分布中抽取。

wilcox.test(y~x,data)
wilcox.test(y1,y2)

# 如果需要多组比较可以采用（各组独立）
kruskal.test(y~A,data)

# 如果需要多组比较可以采用（各组不独立）
friedman.test()
```

7 分布拟合

统计研究中经常面临这样一个问题，获得一组数据，想判断其是否来自某一分布总体。这个总体的概率密度函数为 $f(x,\theta)$。

步骤：（1）假设样本属于某一分布；（2）估计参数；（3）评估拟合效果；（4）拟合优度检验。

```
x.norm<-rnorm(n=200,m=10,sd=2)
hist(x.norm,main="Histogram of observed data")
plot(density(x.norm),main="Density estimate of data")
plot(ecdf(x.norm),main="Empirical cumulative distribution function")
z.norm<-(x.norm-mean(x.norm))/sd(x.norm) ## standardized data
qqnorm(z.norm) ## drawing the QQplot
#qqplot() for any kind of distribution.
abline(0,1) ## drawing a 45-degree reference line


#Weibull distribution

## sampling from a Weibull distribution with parameters
## shape=2.1 and scale=1.1

x.wei<-rweibull(n=200,shape=2.1,scale=1.1)

## theorical quantiles from a Weibull population with
##known paramters shape=2 e scale=1

x.teo<-rweibull(n=200,shape=2, scale=1)
## QQ-plot

qqplot(x.teo,x.wei,main="QQ-plot distr. Weibull")
## a 45-degree reference line is plotted

abline(0,1)
```