R in action读书笔记（7）-第七章：基本统计分析（下） - jpld

7.3相关

相关系数可以用来描述定量变量之间的关系。相关系数的符号（±）表明关系的方向（正相关或负相关），其值的大小表示关系的强弱程度（完全不相关时为0，完全相关时为1）。除了基础安装以外，我们还将使用psych和ggm包。

7.3.1 相关的类型

1.Pearson、Spearman和Kendall相关

Pearson积差相关系数衡量了两个定量变量之间的线性相关程度。Spearman等级相关系数则衡

量分级定序变量之间的相关程度。Kendall’s Tau相关系数也是一种非参数的等级相关度量。

cor()函数可以计算这三种相关系数，而cov()函数可用来计算协方差。两个函数的参数有

很多，其中与相关系数的计算有关的参数可以简化为：cor(x,use=,method=)

x :矩阵或数据框

use :指定缺失数据的处理方式。可选的方式为all.obs（假设不存在缺失数据——遇到缺失数据时将报

错）、everything（遇到缺失数据时，相关系数的计算结果将被设为missing）、complete.obs

（行删除）以及 pairwise.complete.obs（成对删除，pairwise deletion）

method :指定相关系数的类型。可选类型为pearson、spearman或kendall

```
> states<- state.x77[,1:6]
> cov(states)
           Population Income Illiteracy Life Exp  Murder  HS Grad
Population   19931684 571230    292.868 -407.842 5663.52 -3551.51
Income         571230 377573   -163.702  280.663 -521.89  3076.77
Illiteracy        293   -164      0.372   -0.482    1.58    -3.24
Life Exp         -408    281     -0.482    1.802   -3.87     6.31
Murder           5664   -522      1.582   -3.869   13.63   -14.55
HS Grad         -3552   3077     -3.235    6.313  -14.55    65.24
> cor(states)
           Population Income Illiteracy Life Exp Murder HS Grad
Population     1.0000 0.2082     0.1076  -0.0681 0.3436 -0.0985
Income         0.2082 1.0000    -0.4370   0.3400 -0.2300  0.6199
Illiteracy     0.1076 -0.437     1.0000   -0.5880 0.7030 -0.6572
Life Exp      -0.0681 0.3400    -0.5880   1.0000 -0.7810  0.5822
Murder         0.3436 -0.2300    0.7030   -0.7810 1.0000 -0.4880
HS Grad       -0.0985 0.6200    -0.6570   0.5820 -0.4880  1.0000

> cor(states, method="spearman")
           Population Income Illiteracy Life Exp Murder HS Grad
Population      1.000  0.125      0.313   -0.104  0.346  -0.383
Income          0.125  1.000      0.315    0.324  0.217   0.510
Illiteracy      0.313 -0.315      1.000   -0.555  0.672  -0.655
Life Exp       -0.104  0.324     -0.555    1.000 -0.780   0.524
Murder          0.346 -0.217      0.672   -0.780  1.000  -0.437
HS Grad        -0.383  0.510     -0.655    0.524 -0.437   1.000
```

首个语句计算了方差和协方差，第二个语句则计算了Pearson积差相关系数，而第三个语句计算了Spearman等级相关系数

2. 偏相关

偏相关是指在控制一个或多个定量变量时，另外两个定量变量之间的相互关系。你可以使用ggm包中的pcor()函数计算偏相关系数,函数调用格式为：pcor(u,S)

其中的u是一个数值向量，前两个数值表示要计算相关系数的变量下标，其余的数值为条件变量（即要排除影响的变量）的下标。S为变量的协方差阵。

7.3.2 相关性的显著性检验

可以使用cor.test()函数对单个的Pearson、Spearman和Kendall相关系数进行检验。简化后的使用格式为：cor.test(x,y,alternative-,method=)

其中的x和y为要检验相关性的变量，alternative则用来指定进行双侧检验或单侧检验（取值为"two.side"、"less"或"greater"），而method用以指定要计算的相关类型（"pearson"、

"kendall"或"spearman"）。当研究的假设为总体的相关系数小于0时，请使用alternative=

"less"。在研究的假设为总体的相关系数大于0时，应使用alternative="greater"。在默认情况下，假设为alternative="two.side"（总体相关系数不等于0）

cor.test每次只能检验一种相关关系。psych包中提供的corr.test()函数可以一次做更多事情。
corr.test()函数可以为Pearson、Spearman或Kendall相关计算相关矩阵和显著性水平。

>library(psych)>corr.test(states,use=" complete")
参数use=的取值可为"pairwise"或"complete"（分别表示对缺失值执行成对删除或行删

除）。参数method=的取值可为"pearson"（默认值）、"spearman"或"kendall"。

。在多元正态性的假设下，psych包中的pcor.test()函数①可以用来检验在控制一个或多个额外变量时两个变量之间的条件独立性。使用格式为：pcor.test(r,q,n)

其中的r是由pcor()函数计算得到的偏相关系数，q为要控制的变量数（以数值表示位置），n为

样本大小。psych包中的r.test()函数提供了多种实用的显著性

检验方法。此函数可用来检验：

某种相关系数的显著性；

两个独立相关系数的差异是否显著；

两个基于一个共享变量得到的非独立相关系数的差异是否显著；

两个基于完全不同的变量得到的非独立相关系数的差异是否显著。

7.4 t检验
7.4.1 独立样本的t检验

一个针对两组的独立样本t检验可以用于检验两个总体的均值相等的假设。这里假设两组数据是独立的，并且是从正态总体中抽得。检验的调用格式为：t.test(y~x,data)

其中的y是一个数值型变量，x是一个二分变量。调用格式或为：t.test(y1,y2)

其中的y1和y2为数值型向量（即各组的结果变量）。可选参数data的取值为一个包含了这些

变量的矩阵或数据框。可以添加一个参数alternative="less"或alternative="greater"来进行有方向的检验。

> t.test(Prob~So,data=UScrime)

Welch Two Sample t-test

data: Prob by So

t = -3.8954, df = 24.925, p-value = 0.0006506

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-0.03852569 -0.01187439

sample estimates:

mean in group 0 mean in group 1

0.03851265 0.06371269

7.4.2 非独立样本的t检验

非独立样本的t检验假定组间的差异呈正态分布。

t.test(y1,y2,parired=TRUE)其中的y1和y2为两个非独立组的数值向量

> library(MASS)

> sapply(UScrime[c("U1","U2")],function(x)(c(mean=mean(x),sd=sd(x))))

U1 U2

mean 95.46809 33.97872

sd 18.02878 8.44545

> with(UScrime,t.test(U1,U2,paired=TRUE))

Paired t-test

data: U1 and U2

t = 32.4066, df = 46, p-value < 2.2e-16

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

57.67003 65.30870

sample estimates:

mean of the differences

61.48936

7.5 组间差异的非参数检验

7.5.1两组的比较

若两组数据独立，可以使用Wilcoxon秩和检验来评估观测是否是从相同的概率分布中抽得的

Wilcox.test(y~x,data)其中的y是数值型变量，而x是一个二分变量。调用格式或为：

Wilcox.test(y1,y2)其中的y1和y2为各组的结果变量。可选参数data的取值为一个包含了这些变量的矩阵或数据框。默认进行一个双侧检验。可以添加参数exact来进行精确检验，指定alternative="less"或alternative="greater"进行有方向的检验。

Wilcoxon符号秩检验是非独立样本t检验的一种非参数替代方法。它适用于两组成对数据和

无法保证正态性假设的情境。调用格式与Mann－Whitney U检验完全相同，不过还可以添加参数

paired=TRUE。

> sapply(UScrime[c("U1","U2")],median)

U1 U2

92 34

> with(UScrime,wilcox.test(U1,U2,paired=TRUE))

Wilcoxon signed rank test withcontinuity

correction

data: U1 and U2

V = 1128, p-value = 2.464e-09

alternative hypothesis: true location shift is not equal to 0

7.5.2 多于两组的比较

如果各组独立，则Kruskal—Wallis检验将是一种实用的方法。如果各组不独立（如重复测量设计或随机区组设计），那么Friedman检验会更合适。Kruskal－Wallis检验的调用格式为：

Kruskal.test(y~A,data)其中的y是一个数值型结果变量，A是一个拥有两个或更多水平的分组变量（groupingvariable）。（若有两个水平，则它与Mann－Whitney U检验等价。）而Friedman检验的调用格式为：friedman.test(y~A|B,data)

其中的y是数值型结果变量，A是一个分组变量，而B是一个用以认定匹配观测的区组变量（blocking

variable）。

```
> states<-as.data.frame(cbind(state.region,state.x77))

> kruskal.test(Illiteracy~state.region,data=states)

Kruskal-Wallis rank sum test

data: Illiteracy by state.region

Kruskal-Wallis chi-squared = 22.6723, df = 3,

p-value = 4.726e-05
```

欢迎关注我的微信平台