代码清单7-16　使用table2flat()函数转换已发表的数据

```
> treatment <- rep(c("Placebo", "Treated"), times=3)
> improved <- rep(c("None", "Some", "Marked"), each=2)
> Freq <- c(29,13,7,17,7,21)
> mytable <- as.data.frame(cbind(treatment, improved, Freq))
> mydata <- table2flat(mytable)
> head(mydata)
  treatment improved
1   Placebo     None
2   Placebo     None
3   Placebo     None
4   Treated     None
5   Placebo     Some
6   Placebo     Some
[12 more rows go here]
```

对列联表的讨论暂时到此为止，我们将在第11章和第15章中探讨更多高级话题。下面，我们将开始关注各种类型的相关系数。

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）。

本节中，我们将关注多种相关系数和相关性的显著性检验。我们将使用R基础安装中的state.x77数据集，它提供了美国50个州在1977年的人口、收入、文盲率、预期寿命、谋杀率和高中毕业率数据。数据集中还收录了气温和土地面积数据，但为了节约空间，这里将其丢弃。你可以使用help(state.x77)了解数据集的更多信息。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

R可以计算多种相关系数，包括Pearson相关系数、Spearman相关系数、Kendall相关系数、偏相关系数、多分格（polychoric）相关系数和多系列（polyserial）相关系数。下面让我们依次理解这些相关系数。

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= )
```

这些参数详述于表7-3中。

表7-3 cor和cov的参数

参 数	描 述
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

默认参数为use="everything"和method="pearson"。你可以在代码清单7-17中看到一个示例。

代码清单7-17 协方差和相关系数

```
> 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  0.208      0.108   -0.068  0.344 -0.0985
Income         0.2082  1.000     -0.437    0.340 -0.230  0.6199
Illiteracy     0.1076 -0.437      1.000   -0.588  0.703 -0.6572
Life Exp      -0.0681  0.340     -0.588    1.000 -0.781  0.5822
Murder         0.3436 -0.230      0.703   -0.781  1.000 -0.4880
HS Grad       -0.0985  0.620     -0.657    0.582 -0.488  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。举例来说，我们可以看到收入和高中毕业率之间存在很强的正相关，而文盲率和预期寿命之间存在很强的负相关。

请注意，在默认情况下得到的结果是一个方阵（所有变量之间两两计算相关）。你同样可以计算非方形的相关矩阵。观察以下示例：

```
> x <- states[,c("Population", "Income", "Illiteracy", "HS Grad")]
> y <- states[,c("Life Exp", "Murder")]
> cor(x,y)
```

```
           Life Exp Murder
Population   -0.068  0.344
Income        0.340 -0.230
Illiteracy   -0.588  0.703
HS Grad       0.582 -0.488
```

当你对某一组变量与另外一组变量之间的关系感兴趣时，cor()函数的这种用法是非常实用的。注意，上述结果并未指明相关系数是否显著不为0（即，根据样本数据是否有足够的证据得出总体相关系数不为0的结论）。由于这个原因，你需要对相关系数进行显著性检验（在7.3.2节中阐述）。

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。ggm包没有被默认安装，在第一次使用之前需要先进行安装。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。这个示例有助于阐明用法：

```
> library(ggm)
> # 在控制了收入、文盲率和高中毕业率时
> # 人口和谋杀率的偏相关系数
> pcor(c(1,5,2,3,6), cov(states))
[1] 0.346
```

本例中，在控制了收入、文盲率和高中毕业率的影响时，人口和谋杀率之间的相关系数为0.346。偏相关系数常用于社会科学的研究中。

3. 其他类型的相关

polycor包中的hetcor()函数可以计算一种混合的相关矩阵，其中包括数值型变量的Pearson积差相关系数、数值型变量和有序变量之间的多系列相关系数、有序变量之间的多分格相关系数以及二分变量之间的四分相关系数。多系列、多分格和四分相关系数都假设有序变量或二分变量由潜在的正态分布导出。请参考此程序包所附文档以了解更多。

7.3.2　相关性的显著性检验

在计算好相关系数以后，如何对它们进行统计显著性检验呢？常用的原假设为变量间不相关（即总体的相关系数为0）。你可以使用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）。参考代码清单7-18中的示例。

代码清单7-18 检验某种相关系数的显著性

```
> cor.test(states[,3], states[,5])

        Pearson's product-moment correlation
data:  states[, 3] and states[, 5]
t = 6.85, df = 48, p-value = 1.258e-08
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.528 0.821
sample estimates:
  cor
0.703
```

这段代码检验了预期寿命和谋杀率的Pearson相关系数为0的原假设。假设总体的相关度为0，则预计在一千万次中只会有少于一次的机会见到0.703这样大的样本相关度（即p = 1.258e–08）。由于这种情况几乎不可能发生，所以你可以拒绝原假设，从而支持了要研究的猜想，即预期寿命和谋杀率之间的总体相关度不为0。

遗憾的是，cor.test每次只能检验一种相关关系。但幸运的是，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。代码清单7-19中给出了一个示例。

代码清单7-19 通过corr.test计算相关矩阵并进行显著性检验

```
> library(psych)
> corr.test(states, use="complete")

Call:corr.test(x = states, use = "complete")
Correlation matrix
           Population Income Illiteracy Life Exp Murder HS Grad
Population       1.00   0.21       0.11    -0.07   0.34   -0.10
Income           0.21   1.00      -0.44     0.34  -0.23    0.62
Illiteracy       0.11  -0.44       1.00    -0.59   0.70   -0.66
Life Exp        -0.07   0.34      -0.59     1.00  -0.78    0.58
Murder           0.34  -0.23       0.70    -0.78   1.00   -0.49
HS Grad         -0.10   0.62      -0.66     0.58  -0.49    1.00

Sample Size
[1] 50
Probability value
           Population Income Illiteracy Life Exp Murder HS Grad
Population       0.00   0.15       0.46     0.64   0.01     0.5
Income           0.15   0.00       0.00     0.02   0.11     0.0
Illiteracy       0.46   0.00       0.00     0.00   0.00     0.0
Life Exp         0.64   0.02       0.00     0.00   0.00     0.0
Murder           0.01   0.11       0.00     0.00   0.00     0.0
HS Grad          0.50   0.00       0.00     0.00   0.00     0.0
```

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"。这里可以看

到，人口数量和高中毕业率的相关系数（–0.10）并不显著地不为0（p = 0.5）。

其他显著性检验

在7.4.1节中，我们关注了偏相关系数。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。

在结束这个话题之前应当指出的是，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。

参阅help(r.test)以了解详情。

7.3.3　相关关系的可视化

以相关系数表示的二元关系可以通过散点图和散点图矩阵进行可视化，而相关图（correlogram）则为以一种有意义的方式比较大量的相关系数提供了一种独特而强大的方法。这些图形将在第11章中详述。

7.4　t 检验

在研究中最常见的行为就是对两个组进行比较。接受某种新药治疗的患者是否较使用某种现有药物的患者表现出了更大程度的改善？某种制造工艺是否较另外一种工艺制造出的不合格品更少？两种教学方法中哪一种更有效？如果你的结果变量是类别型的，那么可以直接使用7.3节中阐述的方法。这里我们将关注结果变量为连续型的组间比较，并假设其呈正态分布。

为了阐明方法，我们将使用MASS包中的UScrime数据集。它包含了1960年美国47个州的刑罚制度对犯罪率影响的信息。我们感兴趣的结果变量为Prob（监禁的概率）、U1（14~24岁年龄段城市男性失业率）和U2（35~39岁年龄段城市男性失业率）。类别型变量So（指示该州是否位于南方的指示变量）将作为分组变量使用。数据的尺度已被原始作者缩放过。（注意，我原本打算将本节命名为“旧南方的罪与罚”，但是最后理智还是战胜了情感。）

7.4.1　独立样本的t检验

如果你在美国的南方犯罪，是否更有可能被判监禁？我们比较的对象是南方和非南方各州，

① 这里可能是作者的笔误，函数*pcor.test*事实上包含于*ggm*包中。——译者注