```
  Runs=as.numeric(order(Data)<=n)
  W=1
  for (j in 1:(length(Runs)-1)){
    if (Runs[j]!=Runs[j+1])
      W=W+1
    }
  E=2*m*n/N+1;V=2*m*n*(2*m*n-N)/(N^2*(N-1))
  res1[i]=as.numeric(abs(W-E)/sqrt(V)>=1.96)  #the critical values
  }
mean(res1)
```

4.2.3 独立性检验

在实际应用中，除了研究数据的异质性，人们还关心数据之间的相依关系。例如，消费者对不同商品的选择是否和所处的年龄段有关，黑人和白人被判死刑的概率是否显著不同等问题。对于这类分类数据之间的相依关系，我们用列联分析独立性检验。而连续性变量之间的相依关系，我们用相关系数来度量。Pearson 相关系数r 用于度量两变量间的线性关系，是经典的参数检验方法。本节重点介绍的 Spearman ρ 相关系数和 Kendall τ 相关系数是度量变量间单调关系的非参数方法。

1、列联分析检验

本节介绍两个分类变量间的独立性检验方法。二维列联表是指对观测个体的两个特征进行分类计算得到的频数分布表。设 n 个样本可按照两个特征 A (r个水平)和 B (c个水平)进行分类，二维列联表为：

	B_1	B_2	$\cdots$	B_c	合计
A_1	n_{11}	n_{12}	$\cdots$	n_{1c}	$n_{1\cdot}$
A_2	n_{21}	n_{22}	$\cdots$	n_{2c}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_r	n_{r1}	n_{r2}	$\cdots$	n_{rc}	$n_{r\cdot}$
合计	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot c}$	n

其中 n_{ij} 表示同时具有 A_i 和 B_j 特征的实际频数。我们用 p_{ij} 表示第 ij格子的理论频率，相应的第 i 行和第 j 列的理论频率为 $p_{i\cdot} = \sum_{j=1}^{c} p_{ij}$ 和 $p_{\cdot j} = \sum_{i=1}^{r} p_{ij}$，且 n_{ij}/n 是 p_{ij} 的极大似然估计。当行变量和列变量独立时，应该有：$p_{ij} = p_{i\cdot}p_{\cdot j}$。则在原假设成立时，应该有 $n_{ij} = nn_{i\cdot}n_{\cdot j}$。因此，分类变量间的独立性检验就可以转换为 $n_{ij} = nn_{i\cdot}n_{\cdot j}$ 的拟合优度检验。其统计量和拟合优度检验的统计量 V 一致，我们也称之为Pearson χ^2 统计量，即

$$Q = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij} - nn_{i\cdot}n_{\cdot j})^2}{nn_{i\cdot}n_{\cdot j}}.$$

且在原假设下，可以证明 $Q \xrightarrow{L} \chi^2((r-1)(c-1))$。

设有X，Y二个离散型随机变量，分别取r个值和c个值，对应地有二维列联表$r \times c$，作n次观测，在(i,j)格的观测频数为$n_{ij}, i = 1,2,...,r; j = 1,2,...,c.$ 观测值落入(i,j)格的概率为p_{ij}，观测频数服从多项分布，其概率密度为

$$\frac{n!}{\prod_{i=1}^{r}\prod_{j=1}^{c} n_{ij}!} \cdot \prod_{i=1}^{r}\prod_{j=1}^{c} p_{ij}^{n_{ij}}$$

由于$\sum_{i=1}^{r}\sum_{j=1}^{c} p_{ij} = 1$，所以参数空间的独立参数个数为$r \cdot c - 1$，$p_{ij}$的$MLE$ 为$\hat{p}_{ij} = n_{ij}/n$

考虑列联表的独立性检验，原假设$H_0 : p_{ij} = p_{i\cdot} \cdot p_{\cdot j}, i = 1,2,...,r; j = 1,2,...,c$，其中$p_{i\cdot}$ 和$p_{\cdot j}$ 分别是X 和Y 的边缘分布。原假设成立时，$p_{ij} = p_{i\cdot} \cdot p_{\cdot j}$,所以参数空间的独立参数个数为$r + c - 2$个。

这时，p_{ij} 的MLE为$\hat{p_{i\cdot}} \cdot \hat{p_{\cdot j}} = (n_i/n) \cdot (n_j/n)$

该检验问题的似然比统计量为：

$$\Lambda(X) = \frac{\prod_{i=1}^{r}\prod_{j=1}^{c}(\frac{n_{ij}}{n})^{n_{ij}}}{\prod_{i=1}^{r}\prod_{j=1}^{c}(\frac{n_{i\cdot}}{n} \cdot \frac{n_{\cdot j}}{n})^{n_{ij}}}$$

由于$(r \cdot c - 1) - (r + c - 2) = (r-1)(c-1)$，故在原假设成立时，$2ln\Lambda$的极限分布为$\chi^2((r-1)(c-1))$。在$2ln\Lambda \geq \chi^2_{1-\alpha}((r-1)(c-1))$时，拒绝原假设。

思考题：说明对于原假设是正态分布$N(0,1)$ 时，如何用这样的方法去构造检验统计量？

根据Pearson χ^2 统计量的导出思想，我们知道 Q 通过每个小格子中联合频数和边际频数之积的差异来检验分类变量间的独立性。对于上述问题，当

数据来自连续性分布，比如正态分布时，我们可以通过划分区间将连续型变量离散化从而转化为分类变量的独立性检验问题。而这一做法在实际数据分析中也是合理的，例如我们研究年龄对不同商品的选择是否有影响时，我们可以将年龄划分为k个区间，从而研究不同年龄段和商品的选择是否独立。接下来，我们通过一个例子来详细阐述这一做法。取样本量 $n = 100$，假设数据 $(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$ 来自二维正态分布，X 和 Y 的均值均是 0，方差都为 1，相关系数为0.6。我们将 $\{X_1, X_2, \cdots, X_n\}$ 和 $\{Y_1, Y_2, \cdots, Y_n\}$分别划分 $k = 8$ 个区间，然后验证Pearson χ^2 统计量 Q 对该检验问题的功效。

```
###独立检验--列联表检验
library(MASS)
k=8
n=100
Sigma <- matrix(c(1,0.6,0.6,1),2,2)
res1 <- NULL

for (i in 1:1000) {
  data=mvrnorm(n, rep(0, 2), Sigma)
  X_data=data[,1]
  Y_data=data[,2]
  Sx=sort(X_data)
  Sy=sort(Y_data)
  inter_x=seq(min(X_data),max(X_data),by=(max(X_data)-min(X_data))/k)
  inter_y=seq(min(Y_data),max(Y_data),by=(max(Y_data)-min(Y_data))/k)
  left_inter_x=rep(1,n)%*%t(inter_x[1:k]) ##t()表示转置
  right_inter_x=rep(1,n)%*%t(inter_x[2:(k+1)])
  left_inter_y=rep(1,n)%*%t(inter_y[1:k]) ##t()表示转置
  right_inter_y=rep(1,n)%*%t(inter_y[2:(k+1)])

  Data1=(X_data)%*%t(rep(1,k))
  Data2=(Y_data)%*%t(rep(1,k))
```

```
  frequx=(left_inter_x<=Data1)&(Data1<right_inter_x)
  frequy=(left_inter_y<=Data2)&(Data2<right_inter_y)
  frequxy <- t(frequx)%*%frequy

  pi=apply(t(frequxy),2,sum)/n
  pj=apply(frequxy,2,sum)/n

  pij=pi%*%t(pj)
  A=(frequxy-n*pij)^2/(n*pij)
  A[is.nan(A)]=0
  stat1=sum(A)
  res1[i]=stat1>qchisq(0.95,(k-1)^2)

}
mean(res1)
```

1、相关系数检验

事实上，对于两个连续性变量而言，用相关系数来衡量二者间的相关程度更为合适，特别是对于单调关系。本小节主要介绍三种相关系数的定义及其性质，并给出具体的模拟验证三种相关系数的检验功效。

首先考虑两个变量间的线性相关关系。我们知道，变量 X 和 Y 间的相关系数定义为 ρ，即

$$\rho = \frac{E(XY) - E(X)E(Y)}{\sqrt{Var(X)Var(Y)}}.$$

显然，ρ 取值在-1到1之间。$|\rho| = 1$ 说明 X 和 Y 之间存在完全线性关系。如果 X 和 Y 相互独立，则有 $\rho = 0$，反之不然。我们关心两个变量间是否存在简单线性关系，即检验：$H_0 : \rho = 0, H_1 : \rho \neq 0$。

样本 $(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$ 的相关程度用Pearson 相关系数 r 度量，即

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}.$$

当 (X,Y) 服从二元正态分布时，在原假设下，可以证明：

$$t = r\sqrt{\frac{n-2}{1-r^2}} \sim t(n-2).$$

注意到，当数据来自正态分布时，独立和不相关等价。因此，在正态性假设下，可以用统计量 t 检验数据间的独立性。但在实际分析中，很多数据难以满足正态性假定，同时用简单相关系数也不能反映出变量 X 和 Y 间的相依关系。因此，对于检验问题：H_0：X 和 Y 是不相关的，H_1：X 和 Y 是相关的，Pearson 相关系数 r 不再适用。Spearman r_s秩相关系数和Kendall τ 相关系数采用非参数方法，能够度量数据间的单调关系。

首先介绍Spearman r_s 秩相关系数。记 (X_i,Y_i) 在各自样本中的秩为 (R_i,S_i)，如果X 和 Y 具有一定的单调关系，那么X 和 Y 的秩应该具有一定的线性相关关系。因此，Spearman r_s秩相关系数通过研究两组数据秩的线性相关系数检验原始数据的单调性，定义为

$$r_s = \frac{\sum_{i=1}^{n}(R_i-\bar{R})(S_i-\bar{S})}{\sqrt{\sum_{i=1}^{n}(R_i-\bar{R})^2\sum_{i=1}^{n}(S_i-\bar{S})^2}}. \tag{4.2.2}$$

其中，$\bar{R}=\sum_{i=1}^{n}R_i/n$ 和 $\bar{S}=\sum_{i=1}^{n}S_i/n$ 是样本的平均秩，且 $\bar{R}=\bar{S}=(n+1)/2$。另一方面，用 $d_i^2=(R_i-S_i)^2$ 表示个体 X_i 和 Y_i 在各自维度的秩差。如果 $d_i^2, i=1,2,\cdots,n$ 一致的大，说明两组数据间存在负相关关系，如果 $d_i^2, i=1,2,\cdots,n$ 一致的小，说明两组数据间存在正相关关系。因此，也可以用 $\sum_{i=1}^{n}d_i^2$ 度量两组数据间的相关关系。简单的计算表明，(4.2.2)式也可以简化为

$$r_s = 1-\frac{6\sum_{i=1}^{n}d_i^2}{n(n^2-1)}.$$

和Pearson 相关系数 r 一样，$-1\le r_s\le 1$。当样本量 n 充分大时，在原假设下，可以证明统计量 r_s 的渐近正态性，即

$$r_s\sqrt{n-1}\xrightarrow{L}N(0,1).$$

Kendall τ 相关系数的思想是用所有 X 和 Y 的数对中方向一致和方向相反的数对比例差异来检验 X 和 Y 的单调关系。具体地，n 个观测

数据 $(X_1,Y_1),(X_2,Y_2),\cdots,(X_n,Y_n)$ 中可以产生 $\binom{n}{2}$组数对，即 $\{X_i,X_j\}$ 和 $\{Y_i,Y_j\}$，$1\le i\le j\le n$。如果 $(X_i-X_j)(Y_i-Y_j)>0$，则称该数对方向一致，如果 $(X_i-X_j)(Y_i-Y_j)<0$，则称该数对方向相反。当数据呈正相关时，方向一致的数对倾向较多；当数据呈负相关时，方向相反的数对倾向较多。因此，两种数对的数目之差可以用来衡量两变量间的相关程度。记 $sign(\cdot)$ 为符号函数，Kendall τ 相关系数定义为：

$$\tau=\binom{n}{2}^{-1}\sum_{1\le i\le j\le n}sign\{(X_i-X_j)(Y_i-Y_j)\}.$$

当所有数对的方向均一致时，$\tau=1$，当所有数对的方向均相反时，$\tau=-1$。因此，和前两种相关系数一致，Kendall τ 相关系数也满足 $-1\le\tau\le 1$。当样本量 n 充分大时，在原假设下，可以证明统计量 τ 的渐近正态性，即

$$\tau\sqrt{\frac{9n(n-1)}{2(2n+5)}}\xrightarrow{L}N(0,1).$$

接下来，我们通过一个模拟来验证Spearman r_s 秩相关系数和Kendall τ 相关系数的检验功效。取样本量$n=100$，按照以下设置产生数据，$X_i\sim N(0,1)$，$Y_i=\beta_1X_i+\beta_2X_i^2+\epsilon_i$，$i=1,2,\cdots,n$，其中 $\beta_1=0.5$，$\beta_2=-1$，$\epsilon_i\sim N(0,1/n)$。模拟代码如下：

```
n=100
beta1=0.5
beta2=-1
res1 <- NULL
res2 <- NULL
for (i in 1:1000){
  X_data=rnorm(n,0,1)
  epsilon=rnorm(n,0,1/sqrt(n))
  Y_data=beta1*X_data+beta2*X_data^2+epsilon
  R=rank(X_data)
  S=rank(Y_data)
  r_s=1-6*mean((R-S)^2)/(n^2-1)
  XX=X_data%*%t(rep(1,n))-rep(1,n)%*%t(X_data)
```

```
  YY=Y_data%*%t(rep(1,n))-rep(1,n)%*%t(Y_data)
  tau=sum(sign(XX*YY))/(n*(n-1))
  res1[i]=abs(r_s*sqrt(n-1))>1.96
  res2[i]=abs(tau*sqrt(9*n*(n-1)/(4*n+10)))>1.96
}
c(mean(res1),mean(res2))
```