非参第 5 次个人作业

薛修梅 2015201589 统计学院

要求编写自定义函数计算W^+的精确分布（大样本下用正态近似），进行 Wilcoxon 秩检验。

W 的精确分布:

```
accu_wilcox<-function(x,test_value,alternative=c("two.sided","less","greater"))
{
  x_factor<-as.factor(x)                    # 将 x 转化为因子
  alternative<-match.arg(alternative)    # 参数匹配,默认"two.sided"
  n<-length(x)                              # 样本量

  k1<-vector()
  k2<-vector()
  k3<-vector()
  k4<-vector() # 存放秩, 用来计算 W 的值
  for(i in 1:n)
  {
    k1[i]=abs(x[i]-test_value)
  }
  k2<-rank(k1) # 对取绝对值后数据求秩
  for(i in 1:n)
  {
    if((x[i]-test_value)>0){k3[i]=k2[i]}
    else{k3[i]=0}
  }
  W1=sum(k3)   # W+
  for(i in 1:n)
  {
    if((x[i]-test_value)<0){k4[i]=k2[i]}
    else{k4[i]=0}
  }
  W2=sum(k4)   # W-
  if(alternative=="two.sided"){W=min(W1,W2)}
  if(alternative=="greater"){W=W2}
  if(alternative=="less"){W=W1}

  a=c(1,1) # 参考 P57:已知 W, 求对应概率
  for(i in 2:n)
  {
```

```
    t=c(rep(0,i),a)
    a=c(a,rep(0,i))+t
    p=a/(2^i)
  }

  p1<-sum(p[1:W]) # 计算分布密度（即累积概率）
  # 计算 p_value
  if(alternative=="two.sided")
  {
    if(p1>0.5){p_value<-(1-p1)*2}else{p_value<-p1*2}
  }else if(alternative=="less"){p_value<-p1}else if(alternative=="
greater"){p_value<-1-p1}

  DNAME <-deparse(substitute(x)) # 数据名
  ## substitute(),替换表达式中的变量
  # 如果我们希望在表达式中使用变量并且希望这些变量在运行过程中做出相
应改变，就可以使用 substitute 函数
  ## deparse(),把表达式逆解析为字符
  METHOD <- "Wilcoxon Test With Accurate Distribution" # 标题
  structure(list(alternative=alternative,p.value=p_value,method=M
ETHOD,data.name=DNAME),class="htest") # htest,模板名
}
```

大样本下用正态近似求分布：

```
normappr_wilcox<-function(x,test_value,alternative=c("two.sided",
"less","greater"))
{
  x_factor<-as.factor(x)                          # 将 x 转化为因子
  alternative<-match.arg(alternative)    # 参数匹配,默认”two.sided”
  n<-length(x)                                    # 样本量

  k1<-vector()
  k2<-vector()
  k3<-vector()
  k4<-vector() # 存放秩，用来计算 W 的值
  for(i in 1:n)
  {
    k1[i]=abs(x[i]-test_value)
  }
  k2<-rank(k1) # 对取绝对值后数据求秩
  for(i in 1:n)
  {
    if((x[i]-test_value)>0){k3[i]=k2[i]}
    else{k3[i]=0}
```

```
  }
  W1=sum(k3)   # W+
  for(i in 1:n)
  {
    if((x[i]-test_value)<0){k4[i]=k2[i]}
    else{k4[i]=0}
  }
  W2=sum(k4)   # W-
  if(alternative=="two.sided"){W=min(W1,W2)}
  if(alternative=="greater"){W=W2}
  if(alternative=="less"){W=W1}

  l=vector()     # 求 k2 各结长
  k5=sort(k1)    # 返回排序后的向量
  j=2            # 初始值为 2
  l[1]=sum(k5==k5[1]) # 和第一个数相同的个数
  for(i in 1:(n-1))
  {
    if(k5[i]!=k5[i+1])
    {
      l[j]=sum(k5==k5[i+1]) # 类推
      j=j+1
    }
  }
  m0=vector()    # 中间统计量
  for(i in 1:length(l))
  {
    m0[i]=(l[i]^3-l[i])/48
  }
  m=sum(m0)

  if(n<=30)      # 参考 P58 公式，构造渐进正态统计量 z
  {
    if(W>=(n*(n+1)/4)) {z=(W-n*(n+1)/4+0.5)/(sqrt(n*(n+1)*(2*n+1)/
24)-m)}else {z=(W-n*(n+1)/4-0.5)/(sqrt(n*(n+1)*(2*n+1)/24)-m)}
  }else {z=(W-n*(n+1)/4)/(sqrt(n*(n+1)*(2*n+1)/24)-y)}

  p1=pnorm(z)    # 计算对应的 p_value
  if(alternative=="two.sided")
  {
    if(p1>0.5){p_value<-(1-p1)*2}else{p_value<-p1*2}
  }else if(alternative=="less"){p_value<-p1}else if(alternative=="
greater"){p_value<-1-p1}
```

```
  DNAME <-deparse(substitute(x)) # 数据名
  ## substitute(),替换表达式中的变量
  # 如果我们希望在表达式中使用变量并且希望这些变量在运行过程中做出相
应改变，就可以使用 substitute 函数
  ## deparse(),把表达式逆解析为字符
  METHOD <- "Approximate Normal Wilcoxon Test" # 标题
  structure(list(alternative=alternative,p.value=p_value,method=M
ETHOD,data.name=DNAME),class="htest") # htest,模板名
}
```

例 2.12

由直方图，没有明显迹象表明数据的分布非对称，所以可以采用 Wilcox 秩检验。

下面比较三种函数的检验效果：

```
data<-c(310,350,370,377,389,400,415,425,550,295,325,296,250,340,2
98,365,375,360,385)# 存入数据
# 内置函数 wilcox.test()
wilcox.test(data,mu=320)
```

运行结果：

```
> wilcox.test(data,mu=320)

        Wilcoxon signed rank test

data:  data
V = 158, p-value = 0.009453
alternative hypothesis: true location is not equal to 320
```

```
# 精确分布
accu_wilcox(data,320)
```

运行结果：

```
> accu_wilcox(data,320)

        Wilcoxon Test With Accurate Distribution

data:  data
p-value = 0.008232
alternative hypothesis: two.sided
```

```
# 大样本近似正态
normappr_wilcox(data,320)
```

运行结果：

```
> normappr_wilcox(data,320)

        Approximate Normal Wilcoxon Test
```

```
data:  data
p-value = 0.01061
alternative hypothesis: two.sided
```

注：观察后两个函数的输出结果，感受自定义函数最后的模板设置。

结论 我们可以看到三种检验的 p 值均小于 0.05，故拒绝原假设，即垃圾邮件数量的中心位置不是 320 封。
同时，精确分布的 p 值最小，可以看出它在小样本情况下对数据的信息利用得最为完整；而大样本的正态近似得到的 p 值最大，说明在小样本场合大样本拟合的效果较差，也为我们接下来进行 Wilcox 秩检验提供了方向。

2.1

分析 此题为典型的符号检验。

解 对于假设检验问题：

H_0: 顾客购买的商品平均件数为 10 vs H_1: 顾客购买的商品平均件数不为 10

其中，10 是待检验的平均数值。

(1) 符号检验(略). **p-value = 1.**

(2) 做 Wilcoxon 秩和检验：

```
data<-c(22,9,4,5,1,16,15,26,47,8,31,7)        # 读入数据
# 法一：内置函数 wilcox.test
wilcox.test(data,mu=10,exact = FALSE)        # Wilcoxon 检验
```

运行结果：

```
> wilcox.test(data,mu=10,exact = FALSE)
        Wilcoxon signed rank test with continuity correction

data:  data
V = 53, p-value = 0.2892
alternative hypothesis: true location is not equal to 10
```

注：警告“无法精確計算带连结的 p 值”是因为数据中存在重复的值，一旦去掉重复值，警告就不会出现；可以通过加上 exact=FALSE 的参数解决，不过得到的 P 值是一个近似值。

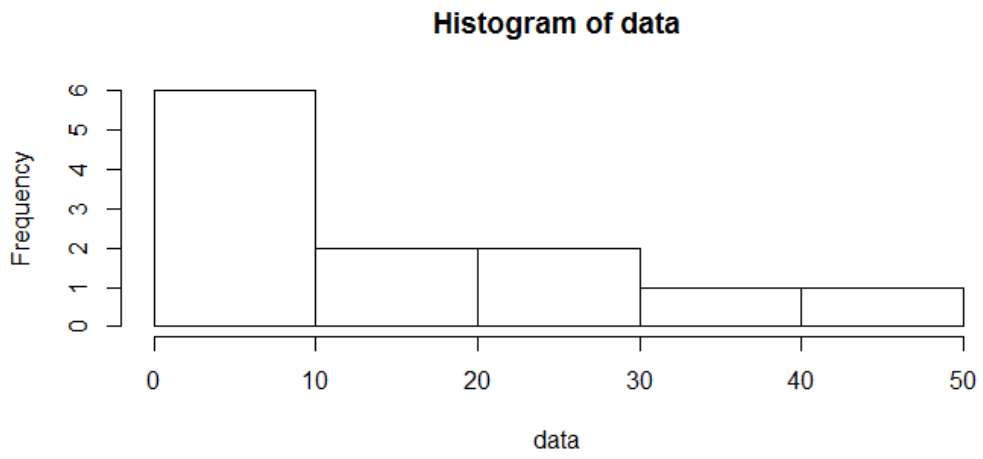
```
# 法二：小样本精确分布
accu_wilcox(data,10)
```

运行结果：

```
> accu_wilcox(data,10)
        Wilcoxon Test With Accurate Distribution

data:  data
p-value = 0.2661
alternative hypothesis: two.sided
```

可以看出，p 值均>0.05，不能拒绝原假设，但 p 值与符号检验（p 值为 1）相比较小。合理推测是因为 wilcoxon 秩检验考虑了数据在待检验均值两侧的分布疏密情况，检验要求更高，故 p 值更小。

检验 作出顾客购买商品件数的直方图：



结论 虽然两个检验的结论相同，但我们认为符号检验可靠。因为后者是基于对称分布做的检验；而由上图可知，有明显迹象表明本题数据是不对称的。所以针对本题数据，wilcoxon.test 不可靠。

2.4

分析 本题考查符号检验在配对样本比较中的应用。

解 题目问“X 和 Y 是否存在显著差异”，设联赛 1 和联赛 2 的三分球得分次数分别为 X 和 Y，设 Z=X-Y，假设检验问题转化为：

$$H_0: E(z) = 0 \quad \text{vs} \quad H_1: E(z) \neq 0$$

(1) **补充**符号检验的 R 代码实现(第 3 次作业仿照例 2.6 直接用 z 统计量求解).

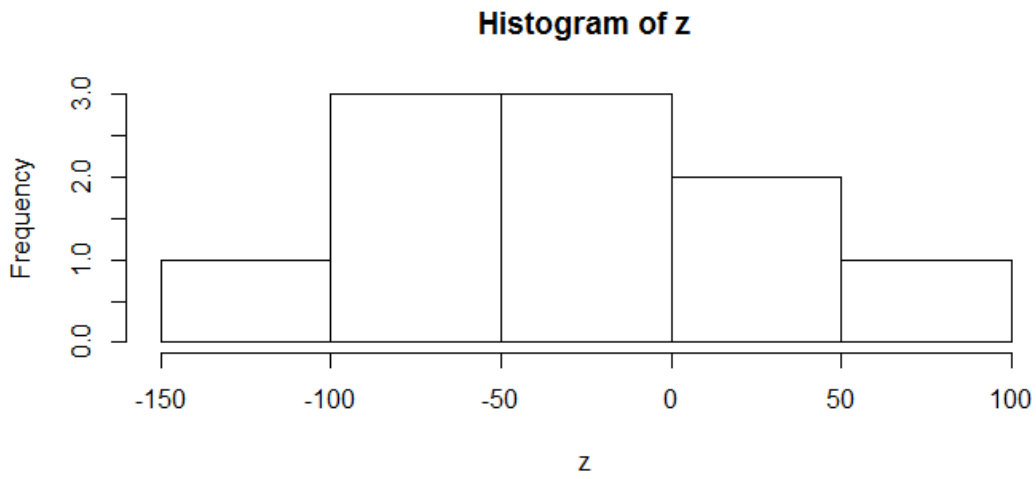
```
x<-c(91,46,108,99,110,105,191,57,34,81)
y<-c(81,51,63,51,46,45,66,64,90,28)        # 读入数据
z<-y-x
sp<-sum(z>0)
sm<-sum(z<0)                                # 计算符号
n<-sp+sm
k<-min(sp,sm)
binom.test(k,n,0.5)                         # 进行符号检验
```

运行结果：

```
> binom.test(k,n,0.5)

        Exact binomial test

data:  k and n
number of successes = 3, number of trials = 10, p-value = 0.3438
alternative hypothesis: true probability of success is not equal to
 0.5
95 percent confidence interval:
 0.06673951 0.65245285
sample estimates:
probability of success
                   0.3
```

可以看出，p 值为 0.3438>0.05，故不能拒绝原假设，即没有充分证据认为两个联赛三分球得分次数存在显著性差异。

(2) 对 z 作直方图如下图所示：



可见 z 的分布不存在显著不对称的迹象，可以做 Wilcoxon 检验。

```
# 法一：内置函数 wilcox.test
wilcox.test(z)      # Wilcoxon 符号秩检验
```

运行结果：

```
> wilcox.test(z)

        Wilcoxon signed rank test

data:  z
V = 10, p-value = 0.08398
alternative hypothesis: true location is not equal to 0
```

```
# 法二：小样本精确分布
accu_wilcox(z,0)          # 精确分布
```

运行结果：

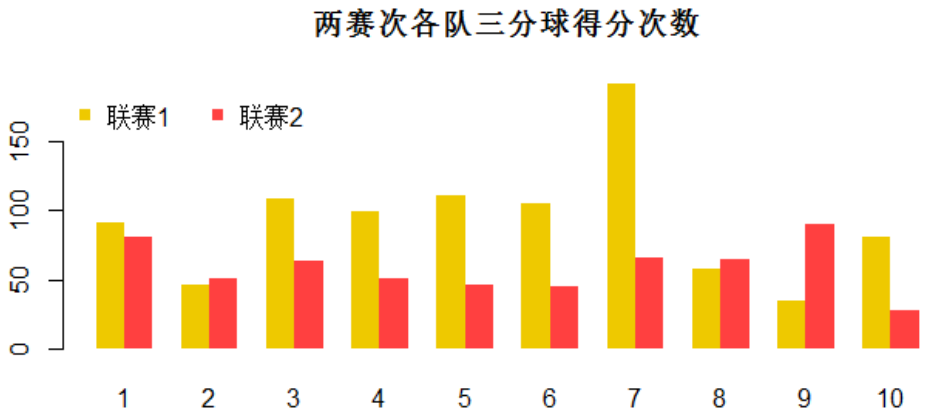
```
> accu_wilcox(z,0)

        Wilcoxon Test With Accurate Distribution

data:  z
p-value = 0.06445
alternative hypothesis: two.sided
```

p 值为均>0.05，同样不能拒绝原假设，即没有充分证据认为两个联赛三分球得分次数存在显著性差异。

(3) 在上述检验过程中，由于数据的分布不存在显著不对称的迹象，所以 wilcox.test 是可靠的，所以 wilcox.test 更好。事实上，wilcox.test 的 P 值相对小了很多，更能区分差异。在检验可靠的情形下，P 值越小越好。
下面进一步验证 p 值变小的原因.

作出在两次篮球联赛中，各球队三分球得分次数的对比柱形图：

```
data<-data.frame(x,y)                                    # 存入数据
barplot(t(as.matrix(data)),beside=TRUE,col=c("gold2","brown1"),axes=TRUE,border=NA,names.arg=c(1:10),main="两赛次各队三分球得分次数")                                                    # 作图
legend("topleft",pch=c(15,15),legend=c("联赛 1","联赛 2"),col=c("gold2","brown1"),bty="n",horiz=TRUE)  # 添加图例
```



可以看出，各球队在两次比赛中的进球次数在数量上有显著差异。若仅仅记正负，不比较绝对值的大小，很容易损失原有数据中的信息，降低检验标准。所以做配对 Wilcoxon 符号秩检验时 p 值更小，效果也更好。

结论 显著性水平 α = 0.05 时，两种方法均证据不足，不能拒绝原假设，即没有充分证据显示两次联赛的三分球得分次数存在差异。但根据本题数据的自身特点，Wilcoxon 秩检验的效果更好。

2.12

分析 检验试验误差是否为正态分布且随机，即对误差项做正态分布一致性检验及符号检验。

解 参照例 2,9，先验证试验误差分布是否随机。

注：由上节课的讨论及证明，例题中的方法存在三个问题使其较为粗陋，改进后检验信度和效度明显提高：
1）默认观测值基于平均水平的误差项与作物品种无关；
2）均值作为点估计是一种不太精细的检验；
3）缺少数据的预处理。
受限于可操作性，在这里我们仅就 3）以±3 σ 为标准做极端值处理（无极端值），特此说明。

假设检验问题：

$$H_0$$: 试验误差分布随机　vs　$$H_1$$: 试验误差分布不随机

由完全随机设计观测值

$$x_{ij} = \hat{\mu} + \hat{\mu}_i + \hat{\varepsilon}_{ij} = \bar{x}_{..} + (\bar{x}_{i.} - \bar{x}_{..}) + (x_{ij} - \bar{x}_{i.})$$. 可知试验误差为 $\varepsilon_{ij} = x_{ij} - \bar{x}_{i.}$，首先计算每个品种的均值 $\bar{A}$=5.3, $\bar{B}$=5.7, $\bar{C}$=5.9, $\bar{D}$=6.3, 再记录各区组实际收成与各自误差成分之间出现顺序为正和负的情况。

符号检验：

```
# 存入各品种水稻在四个区组的表现
a<-c(4.7,5.2,6.2,5.1)
b<-c(5.0,5.4,6.7,5.7)
c<-c(5.7,5.3,6.9,5.7)
d<-c(5.4,6.5,7.4,5.9)
# 计算误差项
a1<-c(4.7,5.2,6.2,5.1)-mean(a)
b1<-c(5.0,5.4,6.7,5.7)-mean(b)
c1<-c(5.7,5.3,6.9,5.7)-mean(c)
d1<-c(5.4,6.5,7.4,5.9)-mean(d)
# 计算误差为正的个数
n1<-length(which(a1>0))+length(which(b1>0))+length(which(c1>0))+l
ength(which(d1>0)) # 符号检验
binom.test(n1,12,0.5)
```

运行结果：

```
> binom.test(n1,12,0.5)

        Exact binomial test
```

```
data:  n1 and 12
number of successes = 5, number of trials = 12, p-value = 0.7744
alternative hypothesis: true probability of success is not equal to
 0.5
95 percent confidence interval:
 0.1516522 0.7233303
sample estimates:
probability of success
           0.4166667
```

可见 p 值为 0.7744 > 0.05，没有充分证据拒绝原假设，即没有充分证据说明试验误差分布不随机。

接着我们对误差项做正态分布一致性检验。

法一 chi-test:

```
x<-cbind(a1,b1,c1,d1)

A<-table(cut(x,br=c(-1.2,-0.6,0,0.6,1.2)))

# cut 将变量区域划分为若干区间,table 计算因子合并后的个数

p<-pnorm(c(-0.6,0,0.6,1.2),mean(x),sd(x))

p<-c(p[1],p[2]-p[1],p[3]-p[2],1-p[3])

chisq.test(A,p=p)
```

运行结果:

```
> chisq.test(A,p=p)

        Chi-squared test for given probabilities

data:  A
X-squared = 5.3735, df = 3, p-value = 0.1464
```

注：运行时出现警告“In chisq.test(A, p = p) : Chi-squared 近似算法有可能不准”，事实上对于连续型变量的优度拟合，卡方检验并不是理想的方法：分组不同，拟合的结果可能不同；并且需要有足够的样本含量。

尝试通过**改变分组方式**验证.

```
A<-table(cut(x,br=c(-1.2,-0.4,0.4,1.2)))
p<-pnorm(c(-0.4,0.4,1.2),mean(x),sd(x))
p<-c(p[1],p[2]-p[1],1-p[2])
chisq.test(A,p=p)
```

运行结果:

```
> chisq.test(A,p=p)
```

```
        Chi-squared test for given probabilities

data:  A
X-squared = 0.14189, df = 2, p-value = 0.9315
```

我们发现 p 值发现了显著变化，说明在对连续型变量进行优度拟合时，要谨慎使用卡方检验。

法二 K-S test:

```
ks.test(jitter(x),pnorm,mean(x),sd(x))
```

运行结果:

```
> ks.test(jitter(x),pnorm,mean(x),sd(x))

        One-sample Kolmogorov-Smirnov test

data:  jitter(x)
D = 0.19583, p-value = 0.5099
alternative hypothesis: two-sided
```

注：当数据中存在重复值时会出现警告“Kolmogorov - Smirnov 检验里不应该有连结”，可以通过做 jitter(x)做小扰动解决。

法三 shapiro test:

```
shapiro.test(x)
```

运行结果:

```
> shapiro.test(x)

        Shapiro-Wilk normality test

data:  x
W = 0.88231, p-value = 0.04216
```

可以看到前两种检验的 p 值均大于 0.05，而 Shapiro 检验的 p 值为 0.04 < 0.05。
考虑到 Shapiro test 有较高的检验效能（相对于其他的正态性检验，如 K-S test 等）且 p 值接近 0.05，而 K-S test 的 p 值为 0.5，因此可以判定数据没有明显背离正态分布。
如果试验的后续目的是进行 t 检验或方差分析等，由于这些方法对数据背离正态分布并不敏感的，操作者仍然可以使用，而不必理会正态分布的问题。

结论 在 α=0.05 时，没有充分证据表明试验误差分布非随机、非正态。

2.14

分析 这是一个独立性检验问题。如果发病率与季节无关，即二者独立，则发病人数在一年四个季节是均匀分布的（发病率为 1/4）；否则两者相关。

解 假设 pi(i=1、2、3、4)为人们第 i 个季节发病的概率，则假设检验问题为：

$$H_0: p1 = p2 = p3 = p4 = \frac{1}{4} \quad \text{vs} \quad H_1: p1、p2、p3、p4 \text{ 不全等}$$

做 Pearson 卡方检验：

```
chisq.test(V) # 默认均匀分布
```

运行结果：

```
> chisq.test(V)
Chi-squared test for given probabilities

data:  V
X-squared = 10.533, df = 3, p-value = 0.01454
```

注：也可通过计算卡方得到 p 值.

```
V<-c(495,503,491,581);p<-1/4;n<-sum(V);df<-4-1;
chi2<-sum((V-n*p)^2/(n*p))        # 卡方
(pvalue<-1-pchisq(chi2,df))       # 得到 p 值
```

结论 在 α=0.05 时拒绝原假设，即认为发病率与季节有关。具体地说，冬天的发病率高(p=0.28)。当然，由于本题数据仅仅来自一个医院，其代表性值得商榷。为了得到更为科学的结论，我们应该规范抽样，确保样本具有代表性。