基于 R 语言的非参数检验研究

叶飞武，齐滨
（河海大学水文水资源学院，南京，210098）

摘要：非参数检验是统计推断的重要组成内容，因其在总体分布未知时的优越性使得该方法得到广泛的运用。而 R 作为一款编程语言灵活、方便，图形功能强大且具有不断更新的软件包的统计软件正在受到许多工作者的热爱。结合上述两者的优点，本文利用 R 软件对太湖流域天生港站和江阴站年日最高潮位进行非参数检验分析，结果表明 R 软件对非参数检验有较好的适用性，两站的潮位相关性较好。

关键词：概率论与数理统计；R；函数；相关
中图分类号：O212.7

Study on Non-parametric Tests Based on R Language

YE Feiwu, Qi Bin
(Institute of Hydrology and Water Resources, Hohai University ,Nanjing,210098)

Abstract: Non-parametric test is the important component of statistical inference and widely used, because the overall superiority when distribution is unknown. The R programming language as a flexible, convenient, and powerful graphics package with constantly updated statistics software is favored by many workers. Combination of these two advantages, we use R software to do non-parametric tests based on the maximum water level of Tianshenggang and Jiangyin, the results showed that the R software is good for non-parametric test, the two stations has higher correlation.

Key words: Statistical Theory and Methodology ; R; Fuction；Correlation

0 引言

在统计分析中许由于人们往往对总体的分布之甚少，也就很难对总体做出正确的假定。由于非参数方法假设条件较少，运算比较简单，不需要太多的数学和统计理论，且易于理解，在总体分布未知时有很大的优越性，且总是比传统检验安全，因此其适用范围较广。R 语言是为统计计算和图形展示而设计的一种编程语言和统计环境。[1]R 软件相对于其他软件，其特色在于有效的数据处理和保存机制和连贯而又完整的数据分析工具，是交互式数据分析的一个非常好工具。本文主要运用 R 软件对太湖流域的潮位进行非参数检验，为潮位相关计算提供一个方法。

目前非参数检验方法较多，主要有符号检验、Wilcoxon 符号秩检验、kolmogorov-Smirnov 检验等等，这些检验中又分为单样本检验（如 t 检验），两样本检验（如 Wilcoxon 秩和检验），多样本检验如（Kruskal-Wallis 检验)。本文主要的检验方法有 Wilcoxon 符号秩检验、秩相关检验、K-S 检验。

1 非参数统计方法

1.1 Wilcoxon 符号秩检验

假定某个总体的中位数为 M_0，如果样本 X_i 中位数 $m = M_0$ 那么我们就接受样本是来

作者简介：叶飞武（1986-），男，硕士研究生，主要研究方向：工程水文. E-mail: yepiao861121@126.com

自某个总体的假设。定义统计量s^+，s^-，其中s^+为正符号（$x_i-m_0>0$）的个数,s^-为负符号（$x_i-m_0<0$）的个数，若$x_i-m_0=0$,则将x_i从观测数据中去掉，相应的样本个数也减少，然后把$|x_i-M_0|$进行排序，得到$|x_i-M_0|$的秩，之后把x_i-m_0的符号加到相应的秩上。设有统计量W^+，W^-，

$$W^+=\sum_{i=1}^{n} rank|x_i-M_0| \quad \text{其中 } x_i-m_0>0 \tag{1}$$

$$W^-=\sum_{i=1}^{n} rank|x_i-M_0| \quad \text{其中 } x_i-m_0<0 \tag{2}$$

令$W=\min(W^+,W^-)$，则W称为 Wilcoxon 符号秩检验统计量，若W太小时应拒绝零假设。

1.2 秩相关检验

秩相关检验是秩检验的一个重要应用，Pearson 相关检验要求数据来自正态分布，本文则不要求所检验的数据来自来自正态分布的总体，秩相关检验包括 Spearman 和 Kendall 相关检验。

设（x_1,y_1）,(x_2,y_2),...(x_n,y_n)来自二元总体的独立样本，检验两者的相关性，以不相关为原假设，$r_1,\cdots r_n$为$x_1,\cdots x_n$产生的统计量，$R_1,\cdots R_n$为$y_1,\cdots y_n$产生的统计量，[2]有

$$\bar{r}=\frac{1}{n}\sum_{i=1}^{n} r_i=\frac{n+1}{2}=\bar{R}=\frac{1}{n}\sum_{i=1}^{n} R_i \tag{3}$$

$$\frac{1}{n}\sum_{i=1}^{n}\left(r_i-\bar{r}\right)^2=\frac{n^2-1}{12}=\frac{1}{n}\sum_{i=1}^{n}\left(R_i-\bar{R}\right)^2 \tag{4}$$

$$r_s=\left[\frac{1}{n}\sum_{i=1}^{n} r_iR_i-\left(\frac{n+1}{2}\right)\right]\Big/\left(\frac{n^2-1}{12}\right) \tag{5}$$

称r_s为 Spearman（斯皮尔曼）相关系数，以r_s的值检验相关性。

对于 Kendall 检验则引进协同概念，令

$$\Psi(X_i,X_j,Y_i,Y_j)=\begin{cases}1,(X_i-X_j)(Y_i-Y_j)>0\\0,(X_i-X_j)(Y_i-Y_j)=0\\-1,(X_i-X_j)(Y_i-Y_j)<0\end{cases} \tag{6}$$

$$\hat{\tau}=\sum_{1\le i\le j\le n}\Psi(X_i,X_j,Y_i,Y_j)=\frac{K}{C_n^2}=\frac{n_d-n_c}{C_n^2} \tag{7}$$

n_c为协同对子数目，n_d为协同对子数目，t为肯达尔相关系数也。显然

$$K\equiv\sum\Psi=n_c-n_d=2n_c-C_n^2 \tag{8}$$

可以证明

$$K=\sum_{1\le i\le j\le n} sign(r_i-r_j)\cdot sign(R_i-R_j) \tag{9}$$

结合(7)和(9)式就可得到t值，做出相关判断。

1.3 Kolmogorov-Smirnov 检验

K-S 检验是用来一个数据总体的累积分布函数是否是已知的理论分布族$F_0(x)$。K-S 检验属于拟合优度检验。设$S_n(x)$为一个n次随机样本观察值的累积概率分布函数，$S_n(x)=i/n$，i是等于或小于x的所有观测结果的数目，$i=1,2,\cdots n$，$F_0(x)$表示一个特定的累计概率分布函数，对于任一x值，$F_0(x)$值代表小于或等于x值的那些预期结果所占的比例。[3]于是，定义$S_n(x)$与$F_0(x)$之间差值，即

$$D=\left|S_n(x)-F_0(x)\right| \tag{10}$$

其中，$S_n(x)$为经验分布函数，$F_0(x)$为理论分布函数，D越小说明拟合程度越高，则有理由认为样本数来自具有该理论分布的总体。K-S 检验利用的是$D=\max\left|S_n(x)-F_0(x)\right|$中那个最大偏差。即；利用统计量

$$D=\max\left|S_n(x)-F_0(x)\right| \tag{11}$$

即可判定。

2 算例

本文利用 R 语言对太湖流域沿江口门天生港站（2006～2007）和江阴站（2007）的日最高潮位值系列分别用上文所述方法进行检验。为了便于分析和计算，把三者的数据分别保存为 tsg2006.txt、tsg2007.txt 和 jy2007.txt。

2.1 数据输入和处理

启动 R 软件，进入主界面，在 RGUI 界面中输入以下语句：

```
# 数据读取
>tsg2006=scan("g:/tsg2006.txt")
>tsg2007=scan("g:/tsg2007.txt")
>jy2007=scan("g:/jy2007.txt")
# Wilcoxon 符号秩检验,原假设为 tsg2006 与 tsg2007 无显著差别
>wilcox.test(tsg2006,tsg2007,alternative="greater",exact = TRUE，correct=FALSE)  # ①
>wilcox.test(tsg2006-tsg2007,alternative="greater",exact = TRUE，correct=FALSE)  # ②
# 秩相关检验-spearman 检验，原假设为 tsg2007 与 jy2007 不相关。
>cor.test(tsg2007, jy2007,method="spearman",conf.level = 0.95)
# 秩相关检验-kendall 检验，原假设为 tsg2007 与 jy2007 不相关。
>cor.test(tsg2007,jy2007,method="kendall",conf.level = 0.95)
# K-S 检验，原假设为 tsg2007 与 jy2007 是同分布的
>ks.test(tsg2007,jy2007)
```

上述三个检验函数中：alternative 表示备择假设，有单侧或双侧检验；conf.level 表示置信度（默认为 0.95）；correct 表示是否采用连续性修正，

p-value 表示拒绝原假设最小显著性水平，p-value >α(置信度)则保留原假设，反之拒绝；exact

表示是否精确计算 p-value 值。

2.2 检验结果与分析

以上语句运行后，在主界面会显示相应的结果，整理后为表 1。

表 1 各类检验方法成果
Tab.1 Results on different test methods

方法	wilcoxon	spearman	kendall	K-S
置信度	5%	5%	5%	5%
p-value	0.9612	2.2E-16	2.2E-16	0.5211
相关系数	—	0.9499	0.8556	—
结果	√	×	×	√

由表 1 可知：在 0.95 的显著水平下，通过 wilconox 检验原假设，即天生港站 2006 和 2007 年的年日最高潮位无显著差别； 拒绝秩和检验的原假设， 即 2007 年天生港站和江阴站的日最高潮位具有相关关系，spearman 相关系数为 0.9499，kendall 相关系数为 0.8556，两者相关系数均大于 0.85，且为正相关；通过 k-s 检验，即 2007 年天生港站和江阴站的日最高潮位具有相同的分布函数。

前文所用的都是两样本的检验，单样本的检验同样适用。对于两样本的差别比较，两样本检验与单样本检验是一致的，如上文中①和②式是等价的。当然，单样本检验能反映出许多独立的信息，如 ks.test[4~5]的单样本检验可以检验样本是否属于特定分布，但要给定具体分布函数名称和必须参数。若要对大量数据进行非检验，则需加入循环和引用检验结果，然后保存相应的结果。以下针对 kendall 秩检验是编写的一段 R 语句。

```
>tsg=read.table("g:/tsg.txt")
>jy=read.table("g:/tsg.txt")
>kpvalue=array(0,dim=(c(n,n)))   #kpvalue：存放双侧检验 p-value
>cortau=array(0,dim=(c(n,n)))   #cortau：存放 kendall 相关系数
>library(package="Kendall")
>for (i in 1:39){for (j in 1:n)
    { kpvalue[j,i]=Kendall(tsg[[i]] jy[[j]]))$sl
    cortau[j,i]=Kendall(tsg[[i]] jy[[j]]))$tau }}
    library(package="MASS")
    write.matrix(kpvalue,"g:/ kpvalue.txt"
    write.matrix(cortau,"g:/ cortau.txt") #保存结果
```

注：n 为检验样本的年份数

在 R 语言中，除了上述所用的检验函数外，还提供了丰富的相似检验函数，从而使结果更加全面和精确。如对于 spearman 秩相关检验可用 spearnoties 函数，而 kendall 秩相关检验则可用 kendall[6~7]函数。前文中的 Ks.test 只能选择一种备选假设输出结果，而利用 package 包 fBasics 中 ks2Test[8]函数则可把四种情况的 p-value 一一列出。此外，在进行非参数检验的时候，常遇到检验样本的打结现象，即成对样本的某些值相等。在 R 语言的计算会给出“在有连结的情况下无法计算精确 P 值”的警告信息，为此 R 中也有对应的修正检验函数解决此类的问题,如表 2 所示，本文不再进行相关介绍和计算。

表 2 修正函数列表

Tab.2 List on modify functions

检验方法	原函数	所属软件包	修正函数
wilcoxon	wilcox.test	coin	wilcox.manyzeros.exact
spearman	cor.test	kendall	spearnoties
		agricolae	correl
kendall	cor.test	corrperm	Kendall
		agricolae	correl
K-S	ks.test	Matching	ks.boot

3 结论

本文给出了采用 R 语言对高潮位的非参数检验的步骤，通过对检验结果的分析表明表明两站用于检验的数据符合两站之间的天然联系，具有较高的可靠性，可为后续计算提供保证，同时为了应对检验的需求和可能出现的问题还对 R 语言的非参数检验函数进行了补充论述。利用 R 进行非参数检验只是 R 中的一个简单的应用。在 R 中还提供了各式各样的函数处理更加复杂的统计学问题，如回归分析和多元统计分析等。由于许多软件包中函数都是个人编写的，所以在实际的运用中，要理解函数之间的差异，从而选择适当的包（函数）处理实际问题。随着多学科的交叉融合，统计学的应用将会深入各个领域，而 R 统计软件将因其独特的优势而受到越来越多的统计使用者的青睐。

[参考文献] (References)

[1] 叶文春.浅谈 R 语言在统计学中应用[J].中共贵州省党校报,2008,4(116):123-124

[2] 薛毅,陈立萍.统计建模与 R 软件[M].北京:清华大学出版社,2006.

[3] 易丹辉.非参数统计-方法与运用[M].北京：中国统计出版社,1995.

[4] Z. W. Birnbaum, Fred. H. Tingey. One-sided confidence contours for probability distribution functions[J].The Annals of Mathematical Statistics,1951,22(4):592-596

[5] William J. Conover. Practical Nonparametric Statistics[M].New York:John Wiley & Sons,1971.

[6] Davison, A.C, Hinkley, D.V. Bootstrap Methods and Their Application[M].London:Cambridge University Press,1997.

[7] Valz, P.D, Thompson, M.E.Exact inference for Kendall's S and Spearman's rho[J].Journal of Computational and Graphical Statistics, 1994(3):459-472

[8] Lehmann E.L. Testing Statistical Hypotheses[M].New York:John Wiley and Sons,1986.