统计数据异常值的识别及R语言实现

王怀亮
（菏泽学院经济系）

摘　要：从统计理论的角度探讨识别统计数据中的异常值，即Grubbs法、dixon法在识别统计数据中的异常值中具有的重要的作用；介绍了在R语言的环境下，程序包outliers的程序语言，并结合具体实例，说明程序包outliers在识别统计数据中异常值的作用。

关键词：R软件；Grubbs法；Dixon法；异常值

Recognition of Statistical Data Outliers and R Language Realization

Wang Huailiang
(Department of Economics, Heze University)

Abstract: From the viewpoint of statistical theory, the recognition of statistical data outliers is explored, which means the important role of Grubbs method and Dixon method in the recognition of statistical data outliers. The paper also introduces the program language of program package outliers under R language environment, and explains the role of program package outliers in the recognition of statistical data outliers combined with specific examples.

Key words: R software; Grubbs method; Dixon method; outlier

0 引言

近年来，随着人们对统计数据关注度的不断提高，对统计数据的质量要求也越来越高，而要很好地保证统计数据质量，其中之一就要关注统计数据中的异常值。所谓异常值，是指一批数据中有部分数据与整体中其他数据相比存在明显不一致，也称为异常数据，或称离群值。异常值的出现可能是由于记录错误引起的，也可能是由于该数据值不属于这个数据集。异常值是影响统计数据质量的一个非常重要的因素，一直以来，我国的统计界以及社会各界均对此问题给予很高的关注。所以，近年来有关异常值的理论探讨一直是个热点。但是目前研究的重点一直放在统计法律制度的健全以及统计工作程序的完善等方面。虽然这两点确实能提高统计数据的质量，但对于已经形成的统计数据，在进行统计分析之前，我们更关注的是统计数据的误差问题，即所提供的统计数据与客观的社会经济现象实际的数量特征之间的差距问题。异常值的存在，使得统计分析的误差大大增大，小则出现差错，大则可能发生事故，甚至可能会导致严重的宏观决策失误。因此，在利用已得数据进行统计分析之前，必须对异常值进行探测和检验。

在统计软件方面，常用的统计软件有SPSS、SAS、STAT、R、S-PLUS等。R软件是一个自由、免费、开源的软件，是一个具有强大统计分析功能和优秀统计制图功能的统计软件，现已是国内外众多统计学者喜爱的数据分析工具。本文文章在基于R语言的基础上，结合具体实例，说明R软件程序包outliers在识别统计数据中异常值的作用.

1 Grubbs法及基于R语言的Grubbs法检验程序举例

1.1 Grubbs法原理

（1）将测量的数据按大小顺序排列。

$x_1,x_2,x_3,\cdots,x_n$

（2）设第i个数据可疑，计算

$$T_{计算}=\frac{\overline{x}\updownarrow x_i}{s}$$

（3）查表

$T_{计算}>T_{表}$则第i个数据为异常值，否则为正常值。

1.2 基于R语言的Grubbs法检验程序

在R软件中，用outliers包中的Grubbs检验可以检验出数据集中的1个或2个异常值，具体命令如下：Grubbs.test(x,type=10,opposite=FALSE,two.sided=FALSE)

其中x是检测数据向量；type=10表示检测一个异常值，type=11表示检测2个分别处于两个端点的异常值，type=20表示检测2个一侧的异常值；two.sided表示双边检验。

1.3 应用举例

例1：在一次调查中，收集数据如下：

8.3、5.5、14.0、7.5、4.7、9.0、6.5、10.2、7.7、6.2

请用Grubbs法判断是否有异常值？如果有，是哪个？

R程序如下：

```
> utils:::menuInstallPkgs()
>local({pkg<-select.list(sort(.packages
```

(all.available= TRUE)),graphics=TRUE)
+if(nchar(pkg))library(pkg, character.only=TRUE)})
>x<-c(8.3,5.5,14.0,7.5,4.7,9.0,6.5,10.2,7.7,6.2)
>grubbs.test(x)
R分析输出结果：
Grubbs test for one outlier
data: x
G = 2.2595, U = 0.3697, p-value = 0.03051
alternative hypothesis: highest value 14 is an outlier
R分析输出结果分析：
因为p=0.03051<0.05，所以可以判断14为这组数据的异常值。当然如果经过实际情况分析，判定14不是异常值，是正常值。而觉得小值有可能是异常值的话，可以输入命令如下：
> utils:::menuInstallPkgs()
>local({pkg<-select.list(sort(.packages (all.available= TRUE)),graphics=TRUE)
+if(nchar(pkg))library(pkg,character. only=TRUE)})
>x<-c(8.3,5.5,14.0,7.5,4.8,9.0,6.5,10.2,7.7,6.2)
> grubbs.test(x,opposite=TRUE)
R分析输出结果：
Grubbs test for one outlier
data: x
G = 1.1797, U = 0.8282, p-value = 1
alternative hypothesis: lowest value 4.7 is an outlier
R分析输出结果分析：
因为p=1>0.05，所以可以判断4.7为这组数据的正常值。

2 dixon法原理及基于R语言的dixon法检验程序举例

2.1 dixon法原理

设数据集为$x_1,x_2,x_3,\cdots,x_n$，则其顺序统计量为设为：$x(1)<x(2)<\cdots<x(n)$。其中x(1)为最小值，$x(n)$为最大值，当顺序统计量$x(i)$是正态分布时，Dixon给出了不同样本数量n时统计量D的计算公式。当显著水平a为0.05或0.01时，Dixon给出了其临界值$D1$-$a(n)$。若某样本的统计量$D>D1$-$a(n)$，则$x(n)$为异常值，如果某样本的统计量$D' > D1$-$a(n)$，则$x(1)$为异常值，否则都为正常值。

2.2 基于R语言的Dixon法检验程序

在R软件中，用outliers包中的Dixon检验可以检验出数据集中的1个或2个异常值，具体命令如下：
dixon.test(x,type=10,opposite=FALSE,two.sided =TRUE)
其中x是检测数据向量；type=10表示检测适用于数据集为3~7个数据，type=11表示检测适用于数据集为8~10个数据，type=21表示检测适用于数据集为11~13个数据，type=2,2表示检测适用于数据集为14个或14个以上数据，；two.sided表示双边检验。

2.3 应用举例

例2：利用例1中的测量数据集，利用Dixon检验判断是否有异常值？如果有，是哪个？
R程序如下：
> utils:::menuInstallPkgs()
>local({pkg<-select.list(sort(.packages(all. available= TRUE)),graphics=TRUE)
+if(nchar(pkg))library(pkg, character.only=TRUE)})
>x<-c(8.3,5.5,14.0,7.5,4.7,9.0,6.5,10.2,7.7,6.2)
> dixon.test(x,type=11)
R分析输出结果：
Dixon test for outliers
data: x
Q = 0.4471, p-value = 0.0380
alternative hypothesis: highest value 14 is an outlier
R分析输出结果分析：
因为p=0.038<0.05，所以可以判断14为这组数据的异常值。当然如果经过实际情况分析，判定14不是异常值，是正常值。而觉得小值有可能是异常值的话，可以输入命令如下：
> utils:::menuInstallPkgs()
>local({pkg<-select.list(sort(.packages(all. available= TRUE)),graphics=TRUE)
+if(nchar(pkg))library(pkg,character. only=TRUE)})
>x<-c(8.3,5.5,14.0,7.5,4.7,9.0,6.5,10.2,7.7,6.2)
>dixon.test(x, type=11，opposite=TRUE)
R分析输出结果：
Dixon test for outliers
data: x
Q = 0.1296, p-value = 0.7763
alternative hypothesis: lowest value 4.7 is an outlier
R分析输出结果分析：
因为p= 0.7763>0.05，所以可以判断4.7为这组数据的正常值。
综上分析，利用R软件程序outliers包来实现数理统计中的Grubbs法、dixon法非常容易实现，也便于根据自己的实际情况调整程序，易学易记，非常直观，所以在以后的数据处理分析中，要多多利用R软件来实现，以提高自己的数据分析能力。

参考文献：

[1] http://cran.r-project.org/web/packages/outliers/ outliers.pdf.
[2] http://cran.r-project.org.
[3] 薛毅，陈立萍. 统计建模与R软件[M]. 北京; 清华大学出版社, 2009.
(下接3页)

[2] 赵玉君,臧运平, 王慧,等.Internet对我国农产品传统交易模式的影响 I [J]. 广东农业科学, 2009(9): 272-274.
[3] 郭利川,张晋娟.绿色壁垒对我国西部地区出口贸易的影响[J].特区经济,2005(1):96.
[4] 闻宏伟, 王锋, 宋吉林.信息技术提升农产品出口企业竞争优势的机制[C].中国农业工程学会2005年学术年会论文集, 2005.

作者简介:
洪梅香, 女, 1981年10月生, 山东菏泽人, 山东省菏泽学院经济系, 讲师, 硕士研究生, 主要研究物流信息与物流成本管理
电话: 05305633779; 18905408966
电子信箱: mumeiheyue@163.com
联系地址: 山东省菏泽牡丹区双河路1021号菏泽市农业局农广校孟宪仁转交 (274015)

基金项目:
菏泽学院校级课题

(上接7页)

[4] 赵清华,赵立春.Dixon,Grubbs和Conchran法检验程序在环境监测中的应用[J].山东环境, 997(3): 15-16.

作者简介:
王怀亮 (1981－), 男, 汉族, 山东曹县人, 菏泽学院经济系助教, 硕士, 主要从事计量经济统计分析。
电话: 15065097727; QQ: 114574325
电子信箱: sdhzhlw@163.com
联系地址: 山东省菏泽市大学路菏泽学院经济系 (274015)

基金项目:
2011年度山东统计科研重点课题 (KT11050)

(上接12页)

网上购物, 用户常采取搜索的方式, 最先接触到的是网页, 如果在搜索和访问恶意网站时就进行拦截, 无疑可以大大消除安全隐患。安装最新金山毒霸后, 访问恶意网站时, 网购保镖就会第一时间拦截。金山毒霸2011 SP6还可以自动拦截假网上购物、假支付网站、假网上银行。

(2) 网购保镖保护网络支付安全。网购防护软件具备网购防火墙功能, 在进入网购页面时, 网购保镖自动提示安全类别, 拦截未知程序, 加固浏览器, 防止木马入侵篡改浏览器, 从而保护网上购物安全。

(3) 有效拦截网购木马。有一些木马, 潜伏在下载站点中, 寄生在正规的软件中, 网购保镖可以拦截通过下载文件、聊天工具接收文件入侵的木马, 最大限度杜绝可能带来的网购木马隐患。

(4) 利用杀毒软件及时查杀网购木马。网购保镖具有较强的拦截和防护功能, 最新金山毒霸还具有较强的网购木马查杀功能。

3 结语

近几年, 网购诈骗与网上购物如影相随, 并且花样不断翻新, 用户个人在现有条件下只有不断了解网购诈骗的手法, 增强网上购物风险防范意识, 提高自身风险防范技术, 才能最大限度地降低网上购物风险, 充分享受广泛、安全、便捷的网上购物。

参考文献:
[1] 李少尉. CA认证刚上路[J]. 电子商务世界, 2006(5): 28-33.
[2] 刘化君. 网络安全技术[M]. 北京:机械工业出版社, 2010.
[3] 袁德月, 乔月圆. 计算机网络安全[M]. 北京: 电子工业出版社, 2007.
[4] 张殿明, 杨辉. 计算机网络安全[M]. 北京: 清华大学出版社, 2010.

作者简介:
吴延亮, 男, 1974.6出生, 菏泽学院经济系, 讲师, 研究方向: 电子商务、会计信息系统、管理信息系统
电话: 13583008215
电子信箱: hzxywyl@163.com
联系地址: 菏泽市大学路2269号 (274015)

统计数据异常值的识别及R语言实现

作者：王怀亮， Wang Huailiang
作者单位：菏泽学院经济系
刊名：电子技术
英文刊名：Electronic Technology
年，卷(期)：2012(5)