软件Ｒ的安装和使用（视窗电脑）

周迈
Department of Statistics, University of Kentucky

Ｒ是一个免费的统计分析软件（ＧＮＵ版权，这一点与 LINUX 相似）。它几乎是ＳＰＬＵＳ的一个克隆。（不要钱的ＳＰＬＵＳ）.几乎所有从Ｒ中学到的都可以在ＳＰＬＵＳ中应用，反之亦然。而ＳＰＬＵＳ是一个很高质量的，普遍使用的统计软件。美国药品检验局曾批准使用 2 个统计软件。ＳＰＬＵＳ是其中一个，另一个是ＳＡＳ。Ｒ有各种版本，可以在ＵＮＩＸ计算机，苹果机和这里要详细讲的视窗电脑机上运行；包括视窗９５，９８，ＭＥ，ＮＴ，2000, XP 中都能运行。现下的最新版是Ｒ１.３.1（9/2001）.

1. 在视窗计算机上安装Ｒ

如果你有Ｒ的光碟，则可省去下载的时间和麻烦。将光碟置于光驱中。用鼠标点击 SetupR.exe 即启动了安装。

如果你没有Ｒ的光碟，则可以从以下地址下载“SetupR.exe”:

http://cran.us.r-project.org/bin/windows/base/

注意这是一个１５ＭＢ大的文件，所以会需要较长时间下载。

如果安装顺利你将会在ＳＴＡＲＴ－ＰＲＯＧＲＡＭＳ里看到Ｒ。用鼠标选择它即开始运行Ｒ了。在Ｒ里你可以键入 q( ) 或用鼠标选ＦＩＬＥ－ＥＸＩＴ来退出Ｒ。此时有一对话框，问你是否需要把Ｒ中运行的历史存下。此时点 No。（以后你得到有用结果时应点 Yes 将结果和历史存下）。

Ｒ还有许多附加的功能包，不用时不调用，可省内存。如果要调用，在 Ｒ 里用鼠标点 Packages 看看有那些已经安装了（要先安装才能调用）。如果已按装了则用鼠标选择即可。

如果要新安装功能包，则先启动Ｒ，在Ｒ里用鼠标点击

Package→Install package from local zipfile

然后在窗口中找到你要安装的 package 的 zip 文件，选择即可。（比如在光碟上就有，或事先下载）。

现在假定你的电脑已经成功地安装上了Ｒ。

2. 第一节课

下面是Ｒ的演示，我们将输出略去了，只有需要你打字的输入在这儿。＃后面的是注，Ｒ不执行的。注意Ｒ是分大小写的（象 Unix,不象 DOS）。

注意>符是Ｒ的 prompt, 表示Ｒ已经准备接受你的指令。

```
>data1 <- c(21,25, 25,18,44,20,25,15,19,20,30)   # 这产生了一个向量叫 data1
>1s() #(是 el-s,不是壹-s)   看看有什么（看到了 data1 了吗？）
>summary(data1)
>stem(data1)
>hist(data1)        直方图
>data2<-rnorm(100)   产生 100 个正态随机数，放在 data2 里
>hist(data2)
```

>data() 看看有那些数据已经装进 R 。
>data(sunspots) 调入数据 sunspots
>sunspots 将 sunspots 在屏幕上显示出来。
>plot(sunspots)
可以在图上打出你的名字、学号：
>text(locator(1),"your name and ID")
此时用鼠标点击图上的某个地方。
>rm(data1) 将 data1 抹去。
>demo(graphics) 看作图演示，要在指令窗里回车数次。
>q() 退出 R

初学者常范的一个错误是自己产生或定义一个东西，起名和现有的东西重名。比如 c, t, 等等。一开始好象不会导至错误，但后来会引起许多混乱。所以起名时要避免重名。如你想给你的数据或函数起名 mydata，则先试试：
>mydata
Error:Object "mydata"not found
这表明没有叫 mydata 的东西，你可以用此名字。

如果重复一个指令，则可用箭头来调出前面用过的指令，还可以修改。象 Unix 的 K-shell.

2. 1 打印及图像存档

R 有可以点击的"打印"菜单，对于图像窗口和指令窗口都有。也可以点击 File →Print 即可。或直接点打印机的图号即可。

还可将图像拷进 Excel 或 World 中（便利与其它文字一起编辑）。先将 R 的图像窗口点击成为 active,然后点击 File→Copy to clip board→as bitmap。再打开（微软）Word 或 Excel, 在那儿点击 Edit→Paste。这样你的图就到了 Word 或 Excel 里了。
如果点击 File→Save as→postscript 则便将图存档成为 postscript 文件, 等等。

2. 2 内存

R 有它自己的内存管理系统。可以用 gc()来看看有多少内存已经占用。不过从 R 1.2.0 版本开始（现在的最新版本为 1.3.0）你不必再担心内存问题。当然如果你的 P C 机内存不足则 R 的运行会很慢。

2.3 彩色作图和图中的数学符号

R 可以产生彩色图。（可能你在 demo(graphics)中见过了）。用 plot(x,col="red")来得到一个红色的图。如要其它颜色，用 colors()来看 600 多种颜色的名称。
你也可以用 points(x,col="white")来抹去刚才得到的红色点，（假定你用 white 作底色）。
其它函数也有不少可用 col=的，包括 lines()等等。
R 优于 Splus 的一点是 R 可以在图中作出数学符号和希腊字母（与 tex 语言功能相近）。以下是一个简单例子：（更多可见 demo()）。

>plot(rnorm(100),type="n")
>text(20,0,expression(theta(mu)),col="blue")
>text(40,0,expression(theta{"2+x"}),col="blue")

3 第二节课

R 要边试边学，以下是一些常用的函数用法，试试看。

看演示：	demo() or demo(graphics)
删除 x：	rm(x)
看你有什么：	ls()
随机产生 9 个从 20 到 40 的整数（无重复）：	sample(20:40,9,replace=FALSE)
随机分组，18 个东西分 3 组：	sample(1:3,18,replace=TRUE)
随机分组，18 个东西分 3 组，每组 6 个	sample(rep(1:3,6),18,replace=FALSE)
查阅 sample()函数的功能，用法：	?sample
计算 data1 的样本均值	mean(data1)
计算 data1 的样本标准差	sd(data1)
计算 data1 的样本方差	var(data1)
计算 data1 的样本中位数	median(data1)
	range(data1)
	boxplot(data1)
计算 5 的阶乘	prod(1:5)
计算从 20 个东西里取 5 个的不同种取法	choose(20,5)
产生 100 个标准正态分布的随机数	rnorm(100)
产生并把连续随机数离散化	table(cut(rnorm(100),8))

R 的另一种用法是将几个乃至几十个指令存档于一个 ASCII 文件（比如叫 mycode），然后在 R 里打

>source("mycode") 或用鼠标点 File→source R code。

对于现成的数据不必重新打字输入，而可用 R 来读。先将数据整理成 ASCII 文件（如用 wordpad），然后在 R 里作如下指令：（假定你的数据在 text.dat 中）（将数据读入并存在 data3 中）

>data3<-read.table("c:/stat/test.dat",header=TRUE)

另外还可用 scan()来读数据，用 write()来输出数据。

R 可以替代几乎所有的统计表格，得到各种概率

如果 Z 是一个二项分布随机变量，N=25,p=0.3，则 P(Z≦5)为

>pbinom(5,25,0.3)

P(Z=5)为

>dbinom(5,25,0.3)

P(Z≧5)为

>1-pbinorm(4,25,0.3) #请注意是 4 而不是 5。

最后 P(5≦Z≦10)为

>1-(1-pbinorm(10,25,0.3)+pbinom(4,25,0.3) 或

>pbinom(10,25,0.3)-pbinorm(4,25,0.3) 或

>sum(dbinom(5:10,25,0.3)

也可以索性打印出一张二项分布的概率表来(N=25,p=0.3)

```
>dbinom(0:25,25,0.3)
```

为了看得更清楚些，你可以试试：

```
>print(dbinom(0:25,25,0.3),print.gap=2)
```

或

```
>print(cbind(0:25,dbinom(0:25,25,0.3)),print.gap=3)
```

如果 Z 是一个标准的正态分布变量，则 P(Z<1)为

```
>pnorm(1)
```

如果要计算非标准的正态概率，则要给出均值和标准差。如果 Z 是均值为 –2，标准差为 3 的正态随机变更，则 P(Z<1)为：

```
>pnorm(1,mean=-2,sd=3)
```

而 P(2<Z<3)为：

```
>pnorm(3,mean=-2,sd=3)-pnorm(2,mean=-2,sd=3)
```

对于超几何分布的概率，可用 dhyper()或 phyper()来计算：

```
------------------
|   f11   |      | 19
------------------                假设左边这个 2 × 2 的表是我们关心的。要
|         |      | 11                    计算 f11 的分布概率。
------------------
14         16
```

则 f11=6 的概率为

```
>dhyper(6,14,16,19)
```

如果用 phyper 则得到 f11≦6 的概率。

对于卡方分布（自由度为 1 的中心分布），它小于 3.84 的概率为

```
>pchisq(3.84,df=1,ncp=0)
```

4. 一些习题

题 1：如果整个母体（所有人们）对于一件事的观点正好是一半一半（赞成 / 反对）。而我们用随机抽样来进行调查。用 R 来算出以下概率：

(a)随机抽样 10 个人，其中 6 人或以上赞成

(b)随机抽样 100 个人，其中 60 人或以上赞成

(c)随机抽样 1000 个人，其中　600 人或以上赞成

(d)随机抽样 2000 个人，其中 1200 人或以上赞成

(e)随机抽样 1500 个人，其中赞成人数在 300 到 600 之间（包括 300 和 600）。

根据上面计算，如果你随机抽样了 2000 人，其中 1200 人赞成。你还相信一半 / 一半（赞成 / 反对）吗？摆出理由。[所有计算均可用 pbinom()完成]

也可用 R 打印出一个小小的正态分布概率表，请与书中的比较。

```
>pnorm(seq(-3.5,3.5,0.5))
```

设 Z 是一正态分布的随机变量。均值为 2，标准差为 4。请计算 3.085<Z<4.226 的概率。

单样本的 T 检验。先将数据存入一个向量（比如）叫 data6

>data6<-c(33.9,52.4,48.6,53.5,43.8)

要检验 Ho:μ=46.5 Ha: μ<46.5（其实只算显著性），则

>t.test(data6,alternative=”less”,mu=46.5)

另外 2 个对立假设是“greater”和“tow.sided”。这函数除了给出显著性外还给出一个 95%置信区间。

t.test()还可以做两样本 t 检验。假定有 2 组数据叫 xbefore 和 xafter。又假定数据是不配对的。则可检验：Ho: μ=0 Ha: μ<0

>t.test(x=xbefore,y=xafter,alternative=”less”,mu=0,paired=FALSE)

如果数据是配对的，则改 paired=FALSE 为 paired=TRUE.

单样本的百分比检验。假设数据为：1000 试验中 600 成功。要检验成功概率是否是 0.5: Ho:p=0.5;Ha:p≠0.5,则

>prop.test(600,n=1000,p=0.5,alternative=”two.sided”)

另外两个对立的假设是“less”和“greater”.这函数也给出 95%的置信区间。

进一步的讯息可用 ?t.test。查询在线手册。

如果需要有 R 的附加功能包，则可以先查一下有那些装上了。

>library() 列出所有装好的附加功能包。(假定你有 ctest.)

>library(ctest) 这样便将 ctest 调进来了。（其中包括 binom.test 函数）

>library(help=ctest) 看看在 ctest 功能包中包括那些函数。（其中有 binom.test）

>binom.test(600,n=1000,p=0.5,alternative=”two.siced”) 利用 binom.test 来做统计检验

>?binom.test

如果 library()中没有你需要的功能包，则要先安装上（请看第一节）。

函数 binom.test 与 prop.test 相近，只不过 prop.test 用的是近似计算，binom.test 是精确计算。不过 prop.test 的功能适用性更强更广。

R 还可以调用现成的 C 程序和 Fortran 程序。不过比较复杂。要先将 C/Fortran 程序 转换成 dll 可执行文件。然后调用。你应该尽量在 R 中完成你的计算。如果实在有调用 C/Fortran 的需要，则请找英文的文件。

R 的一些特点：

1、R(Splus)是向量语言。几乎所有运算都最好向量化。（会比 for 循环句快很多）。
+, -, *, /, ^, …… 等等。要一个向量的一部分，则可用 [] 来表示下标范围。

2、R 的运算都是以函数来完成的。R 有 3000 个以上的函数。exp(x), log(x), sqrt(x), q( )，c(x,y) …… 等都是函数，x 都可以是向量。

3、你自己在 R 中可以很容易定义新的函数（例子见后）.

4、R 的图像功能很强，可以互动作图，直至满意。

5、对于非常大的数据，R 可能不太合适。 (>1 gig)

6、许多附加功能包在 R 和 Splus 中是完全一样的。（如 survival, bootstrap 等等）

７、R 中的随机数产生方法可以自由选择（如果你担心随机数的质量的话）
８、R 不要钱，可以让学生每人一份。可以在家里做计算。或在笔记本电脑上

如果你要定义新的函数，可以先在 R 外用任何编辑软件编写好(ASCII file 或.TXT file)然后在 R 里边用 source()读进来。也可以直接在 R 里边编写。比如先调 R 的函数 mean，修改后变为你自己的函数 junk。

```
>junk<-edit(mean)
```

如要修改你的函数，则可用：

```
>fix(junk)
```

如果你要定义的函数很短，则可直接在 R 中键入，例如

```
>junk<-function(x)  {x/(x+5)}
```

下面是又一个自我定义的函数例子。（给定样本大小，样本均值和标准差，产生假数据）

```
fakedata<-function(size, xbar, sdd){
        if(sdd<=0) stop("sdd must >0")
        if(!is.numeric(xbar)) stop("xbar must be a real number")
        fake1 <-rnorm(size)
        fake2 <-fake1 – mean(fake1)
        fake2*(sdd/sd(fake2))+xbar
}
```

这个函数在以下情况有用：有时习题中只给出样本大小（=50），样本均值（=11.8）和标准差（=0.6），但没有原始数据。如果要做 t 检验，则可以如下做：

```
>mydata <- fakedata(50,11.8,0.6)
>t.test(mydata, mu=12, alternative="less")
```

进一步的阅读只能看英文了。可以先看“An Introduction to R”。此书也是免费的。在 R 里点击 Help，然后选这本书，可以打印出来阅读。R 的指令和输出除了数字／符号外，都是英文，所以学点英文看来是必要的。