Short Tutorial of R

R—a popular language and environment to statistically explore datasets

Download R 3.0.1：

For Windows: http://cran.r-project.org/bin/windows/base/R-3.0.1-win.exe

For Mac: http://cran.r-project.org/bin/macosx/R-3.0.1.pkg

R is an open source with a considerable number of extensions/packages. After installation, type in "search()" to see the basic packages you have in R. R also has a popular IDE(integrated development environment) and text editor called RStudio.

1. R as a Calculator:

```
> 1+2
[1] 3
> 3^2
[1] 9
#Try built-in functions
> exp(2)-log(100)  # Try  "log(10,100)"
[1] 2.783886
# Define a compound function
> sqrt(abs(-2))
[1] 1.414214
```

```
> a<-1
> b=2      # ("=" is the same as "<-")
> (a+b)^2
[1] 9
#Define a function z=f(x,y)
> f<-function(x, y) z<-(y^2-x^2)*pi
>print(f(1,2))
#See what variables you have
>ls()
#Remove a and b in case of duplication
>rm(a,b) # Remove all with "rm(list=ls())"
```

2. Create Vectors in R:

```
> A<-c(2,3,5,7,11)
> B<-seq(100,108, by=2) # How about "by=3"
> B
[1] 100 102 104 106 108
> c(A,B)
 [1]   2   3   5   7  11 100 102 104 106 108
> A+B
[1] 102 105 109 113 119
```

```
> airports<-c("JFK","LGA","EWR","SFO")
> length(airports)
[1] 4
> airports[4]   #How about airports[-4] ?
[1] "SFO"
> airports[1:3]
[1] "JFK" "LGA" "EWR"
> airports[c(2,4)]
[1] "LGA" "SFO"
```

Besides the Vector, R has other data types like matrix and data frame. We will discuss data frame and skip matrix. Some other useful built-in functions: runif() generating random numbers between 0 and 1, max(), min(), range(), and rnorm(). Try them to see what they are.

3. Do Some Basic Statistics on R

- Load in a Local File(A tennis dataset of US Open named "USOpen.csv"):

```
#Choose "File", Click "New script", you can open a window to edit your script.
> getwd()
#Set the working directory for R in order to analysis your data
> setwd("path of data")  #for example, mine is setwd("C:/Users/Tony Tong/Desktop/R_file")
#Install the package called "foreign" in order to read a .csv file.
> install.packages("foreign")  # For .jason files, use "rjason" package: # install.packages("rjason")
> library("foreign")
>tennis<-read.csv("USOpen.csv")# try "dim()"
>tennis<-tennis[,9:34]  try "tennis[9:34]", "tennis[9,34]" and "tennis[9:34,]" to see the differences
>fix(tennis)
#Sometimes, we drop cases with missing values by using the following statement:
#tennis_complete<-tennis[complete.cases(tennis),]
```

- Try Built-in Functions for Statistics:

```
>mean(tennis$ace1) # try median()

>sd(tennis$ace1)

>quantile(tennis$ace1,c(0.25,0.75)) #you can put any percentage here!

#get a frequency table

>table(tennis$ace1,tennis$ace2)

#get a summary table for all varibles

>summary(tennis)
```

- Do regressions:

Question1: Is the number of winners correlated with the number of errors?

```
#Merge the number of winners and errors separately for all matches

> winner<-c(tennis$winner1, tennis$winner2)

> error<-c(tennis$error1,tennis$error2)

#Use lm() to build a simple linear model

> model1<-lm(winner~error)

> summary(model1)#You should see results with parameters.
```

Question2: Which one among factors--the number of aces, the average speed of second serves and the proportion of first serve in--is more related to the win/loss?

```
#Create a vector of results and merge those three independent variables we are interested in.

> result<-rep(1,1000)

>result<-c(result,rep(0,1000))

>ace<-c(tennis$ace1,tennis$ace2)

>av_Second_serve<-c(tennis$avgSecServe1,tennis$avgSecServe2)

>winner<-c(tennis$winner1,tennis$winner2)

#Use glm to create a logistic model

> model2<-glm(result~ace+av_Second_serve+winner)

> summary(model2)
```

4. Basic visualizations:

#Always refer to the following site about color and shape before you start to plot something on R.

#http://www.phaget4.org/R/plot.html

>x = rnorm(100) # rnorm() is used to generate random numbers conforming to a normal distribution

>plot(x)

>x=rnorm(1000)

>hist(x)

#plot() is used to generate a figure, while points() is used to add more.

>plot(winner[1:1000],error[1:1000],pch=20,col="blue")

>points(winner[1001:2000],error[1001:2000],pch=21,col="green")

#Add linear trend lines

>myline.fit1<-lm(winner[1:1000]~error[1:1000])

>abline(myline.fit1)

>myline.fit2<-lm(winner[1001:2000]~error[1001:2000])

>abline(myline.fit2)

#full statement for a plot:

#plot(x,y, xlab="x axis", ylab="y axis", main="my plot", ylim=c(0,20), xlim=c(0,20), pch=15, col="blue")

5. Advanced Visualizations:

#Install the "car" package and see more advanced features for simple plots:

>install.packages("car")+library(car)

>scatterplot(winner~error|result)

#We can see whether there is a correlative relationship between each pair of variables

>pairs(~winner+error+c(tennis$ace1,tennis$ace2)+c(tennis$double1, tennis$double2))

>scatterplot.matrix(~winner+error+c(tennis$ace1,tennis$ace2)+c(tennis$double1, tennis$double2)|result)

#Make density plot by installing the package "hexbin"

>install.packages("hexbin")+library(hexbin)

> plot(hexbin(winner,error,xbin=30))

#Install the package called "rgl" and make an interactive three dimensional plot

>install.packages("rgl")+library(rgl)

> plot3d(winner,error,ace)

6. Practice After Class:

1. Use help() or "??+function name" for help. Learn these very useful built-in functions for data operation: sort(), union(), intersect(),setdiff(), subset().
2. Try to write a "For loop" in R to find all matches played between Serena Williams and Justin Henin.
3. Plot an accumulative distribution for the number of aces for all players.
4. Grasp the usage of while, which and apply statements.
5. Construct a whole logistic model to see which factor(s) is/are important predictor(s) for results.
6. Search on CRAN and learn the differences between NA, NaN, Inf and NULL in R
7. Read the tutorials for some visualization packages upon your interest, the links are as follows:
   "Car" http://cran.r-project.org/web/packages/car/car.pdf
   "Lattice" http://cran.r-project.org/web/packages/lattice/lattice.pdf
   "Scatterplot3D" http://cran.r-project.org/web/packages/scatterplot3d/vignettes/s3d.pdf
   "Rcmdr" http://cran.r-project.org/web/packages/Rcmdr/Rcmdr.pdf
   "Rgl" http://cran.r-project.org/web/packages/rgl/rgl.pdf

7. Tips and Resources:

As a beginner:

Use print() in your code to locate a mistake.

Use typeof(), class() to identify data types and prevent mistakes

Ask and search asked questions on Stack Overflow

Define meaningful names for variables

Get used to read tutorials on CRAN

Useful Resources:

R basic: http://cran.r-project.org/manuals.html

Textbook for R: http://shop.oreilly.com/product/9780596801717.do

Quick R: http://www.statmethods.net/

Data Analysis Examples with R: http://www.ats.ucla.edu/stat/dae/