**R for MATH1530**

R is a very powerful programming language to perform statistical analysis. Here we restrict ourselves to commands that are useful for the topics we learn in any section of MATH1530. **R is free** and students can download it to their own computers at home. R is also available in all the computers of the university and through citrix. Usually there is more than one way to do something with statistical software, we will focus on the most simple way in each case.

Data files used (they are ANSI text files) : pulserate.txt, CHCH2014.txt, drugsurv.txt

available from [**http://faculty.etsu.edu/seier/IntroStatsBioBook.htm**](http://faculty.etsu.edu/seier/IntroStatsBioBook.htm)

**right click on the name of the files and save them in your computer or Z: or Q: drive.**

**STARTING**

1. **Where can I get R from** ? <http://www.r-project.org> there are many free books and manuals available there too. Updated versions are frequently available.
2. **How to make my life easier**?
3. **With File>Change directory indicate where your data files (if any ) are**. In that way you don’t need to specify the drive later when you read files
4. **If you want to keep the commands in a document, save them in a text file (you could use the extension** .txt or .R ) instead of a Word .doc file because sometimes Word changes the font of the quotes and R does not recognize them. **Remember R is case sensitive**

**READING DATA**

1. **How to read data? You create an object (any name you want) and assign there the data but you can input the data in several ways**
2. **Typing data in the session window.** This is practical if we have few observations

**nbschips<-c( 27,15,15,16,16,24, 27,23,26,22,22,18,22,22,20,20,20,24,24,25,30, 27, 20)**

1. **Reading the data from a file with a single column of numbers. As example we will use the pulse rate of 210 students stored in, save the txt file in the directory that you indicated to R that your data were going to be in and read it from R**

**pulse<-scan('pulserate.txt')**

1. **Reading the data from a file with several variables** (you can use simple or double quotes)

**cookies<-read.table("chch2014.dat",header=TRUE)**

**attach(cookies) ## to attach the names of variables to the data**

**names(cookies) ## to look at the name of the variables**

1. **Typing the data in a worksheet.** First create an empty data frame

**mydata<-data.frame()**

then **use Editor>data editor** and type the name to see the worksheet appear, type in the data



**When done, close the data window. To save the data in a file type**

**write.table(mydata,'nameoffile')**

**PLOTTING AND CALCULATING BASIC STATISTICS**

1. **Histograms, boxplots and stem and leaf displays**

**hist(pulse)**

**boxplot(pulse)**

**stem(pulse)**

you can even make them prettier by adding color

**hist(pulse,col='salmon')**

1. **Descriptive statistics**

**mean(pulse)**

**median(pulse)**

**sd(pulse)**

**var(pulse)**

**summary(pulse)**

1. **Comparing groups (cookies data)**

**boxplot(chips~brand)**

**by(chips,brand,summary) ## gets means + five number summary**

**by(chips,brand,sd) ## calculates standard deviations per group**

**by(chips,brand,mean) ## calculates the mean per group**

1. **Scatter plot , correlation & regression (example altitude of residence & red blood cells)**

**x<-c(0,1840,2200,2200,5000,5200,5750,7400,8650,10740,12000,12200,12300,14200,14800,14900,17500)**

 **y<-c(4.93,4.75,5.40,4.65,5.42,6.55,5.99,5.39,5.44,5.82,7.50,5.67, 6.31,7.05,6.46,6.66,7.37)**

**plot(x,y)**

**cor(x,y)**

**lm(y~x)**

**abline(lm(y~x))**

of course we can make the plot prettier controlling the size, type and color of the icons used, the title of the plot etc. When you want to know all the options you have type ‘help(plot)’

1. **Tables and plots for one categorical variable**

**table(brand)**

**pie(table(brand))**

**barplot(table(brand))**

1. **Tables and plots for two categorical variables**

**mydrugs<-read.table('drugsurv.dat',header=TRUE)**

 **attach(mydrugs)**

 **table(GENDER,Marijuana)**

**barplot(table(GENDER,Marijuana))**

**barplot(table(GENDER,Marijuana),beside=TRUE)**

**PROBABILITY DISTRIBUTIONS**

There are 4 magic letters with regard to probability distributions in R : d, p, q ,r

**d** calculates f(x) (or p(x) in the case of discrete distributions)

**p** calculates the cumulative probability

**q** calculates the quantile for a given probability

**r** generates random numbers from a distribution

1. **Normal Distribution**

**dnorm(x,u,s) calculates f(x) for a normal with mean u and standard deviation s**

**pnorm(x,u,s) calculates P(X≤x)**

**qnorm(p,u,s) calculates the value of x such that P(X≤x)=p**

**rnorm(n,u,s) generates n values from the N(u,s) distribution**

1. **Binomial Distribution**

**dbinom(x,n,p) calculates p(x) for a binomial with parameters n and p**

**pbinom(x,n,p) calculates P(X≤x)**

**qbinom(p,n,p) calculates the value of x such that P(X≤x)=p**

**rbinom(m,n,p) generates m values from the B(n,p) distribution**

If you want to create a binomial table, for example, for n=10 and p=0.37 , you just type

**x<-0:10 ## to create the sequence of values from 0 to 10**

**px<-dbinom(x,10,0.37) ## to calculate the probabilities for each value of x**

**mytable<-cbind(x,px) ## to form a table by binding the 2 colums**

**mytable ## to see the table**

**TESTING HYPOTHESES AND CONFIDENCE INTERVALS**

1. **Selecting a random sample**

**k<-1:1000 ## creating a sampling frame for 1000 individuals**

**mysample<-sample(k,20) ## selecting a random sample of size 20**

1. **T-test for mean. Example : Ho: u=25 Ha: u≠25**

**nbschips<-c( 27,15,15,16,16,24, 27,23,26,22,22,18,22,22,20,20,20,24,24,25,30, 27, 20)**

**t.test(nbschips,alternative=c("two.sided"),mu=25)**

1. **Paired T-test (example: pressure tolerated before and after treatment with cherry juice)**

**before<-c(2.3,2.6,2.5,2,2.4,2.4,2.1,2.5,2,2.2)**

**after<-c(4.3,4.6,4.9,3.8,4.3,4.2,4.1,4.0,3.9,4.3)**

**t.test(before,y=after,alternative=c("less"),paired=TRUE)**

1. **Two sample t-test. Ho: u1=u2 Ha: u1 ≠u2**

**cookies<-read.table("chch2014.dat",header=TRUE)**

**attach(cookies) ## to attach the names of variables to the data**

 **t.test(chips~brand, alternative=c("two.sided"),mu=0,paired=FALSE)**

1. **Test for one proportion** Assume that you want to test Ho:p=0.25 vs Ha: p<0.25, and that in a sample of 2466 you find 574 successes

**binom.test(574,2466,p=0.25,alternative= c("less"))**

1. **Test for two proportions Ho: p1=p2 Ha:p1>p2,** data are from the polio vaccine example

**prop.test(x=c(142,52),n=c(200000,200000), alternative= c("greater"))**

1. **Goodness of fit test**

**obs<-c(315,108,101,32) # read observed frequencies**

**prob<-c(9/16,3/16,3/16,1/16) ## model probabilities**

**chisq.test(obs,p=prob)**

1. **Chisquare test of independence or homogeneity (for raw and tabulated data)**

 **chisq.test(table(GENDER,Marijuana)) ## for raw data**

**thetable<-matrix(c(315,108,101,32),nc=2) ## if given a table enter counts by columns**

**chisq.test(thetable)**

1. **Test of normality**

 **shapiro.test(nbschips)**

**qqnorm(nbschips)**

**qqline(nbschips)**

**Note: R code to apply bootstrapping and randomization test available from**

[**http://faculty.etsu.edu/seier/RcommCh3.txt**](http://faculty.etsu.edu/seier/RcommCh3.txt) ***Edith Seier – October 31, 2014***