Practical 3 – Working with data, vectorising and plotting

Basic Plotting

Q1. Plot sin(x) for $0<x<12\pi$

```
curve(sin(x), from = 0, to = 12*pi)

#or
x<-seq(from=0, to=12*pi, length.out=100);
plot(x,sin(x)); lines(x,sin(x))
```

Cystic Fibrosis dataset

In the H drive you will find the file `cystfibr.txt` which contains a set of measurements on a set of individuals with cystic fibrosis.

Q2. Take a look at this file using a text editor like Wordpad or Notepad.

Q3. Read the data into a data frame and attach to the data frame.

```
a = data.frame(read.table(file = ''H:/cystfibr.txt'', header
= TRUE))
attach(a)
```

Q4. Calculate the following

(i) the number of individuals in the dataset; `nrow(a)`
(ii) the number of variables measured on each individual; `ncol(a)`
(iii) the names of the variables measured on each individual; `names(a)`

(iv) the mean, median, standard deviation and range of each of the variables (use the `apply` function);

```
> apply(a, 2, mean)
    age     sex  height  weight     bmp    fev1      rv
frc     tlc   pemax
 14.480   0.440 152.800  38.404  78.280  34.720 255.200
155.400 114.000 109.120


> apply(a, 2, median)
   age    sex height weight    bmp   fev1     rv    frc
tlc  pemax
  14.0    0.0  156.0   37.2   71.0   33.0  225.0  139.0
113.0   95.0
```

```
> apply(a, 2, sd)
       age        sex     height     weight        bmp
fev1         rv        frc        tlc      pemax
 5.0589854  0.5066228 21.5000000 17.8981256 12.0052766
11.1971723 86.0169557 43.7187984 16.9681073 33.4369058

> apply(a, 2, range)
     age sex height weight bmp fev1  rv frc tlc pemax
[1,]   7   0    109   12.9  64   18 158 104  81    65
[2,]  23   1    180   73.8  97   57 449 268 147   195
```

(v) calculate the correlation between each pair of variables? Which pair are the most correlated?

```
> round(cor(a),2)
         age   sex height weight   bmp  fev1    rv   frc
tlc pemax
age     1.00 -0.17   0.93   0.91  0.38  0.29 -0.55 -0.64
-0.47  0.61
sex    -0.17  1.00  -0.17  -0.19 -0.14 -0.53  0.27  0.18
0.02 -0.29
height  0.93 -0.17   1.00   0.92  0.44  0.32 -0.57 -0.62
-0.46  0.60
weight  0.91 -0.19   0.92   1.00  0.67  0.45 -0.62 -0.62
-0.42  0.64
bmp     0.38 -0.14   0.44   0.67  1.00  0.55 -0.58 -0.43
-0.36  0.23
fev1    0.29 -0.53   0.32   0.45  0.55  1.00 -0.67 -0.67
-0.44  0.45
rv     -0.55  0.27  -0.57  -0.62 -0.58 -0.67  1.00  0.91
0.59 -0.32
frc    -0.64  0.18  -0.62  -0.62 -0.43 -0.67  0.91  1.00
0.70 -0.42
tlc    -0.47  0.02  -0.46  -0.42 -0.36 -0.44  0.59  0.70
1.00 -0.18
pemax   0.61 -0.29   0.60   0.64  0.23  0.45 -0.32 -0.42
-0.18  1.00
```

Q5. Create the following data frames and for each one calculate the mean of each of the variables

(i) a new data frame containing just individuals older than 15

```
> b = a[age>15, ]
> apply(b,2, mean)
        age         sex      height      weight
bmp        fev1          rv         frc         tlc
pemax
```

```
 19.1818182   0.3636364 170.5454545  51.8454545
79.4545455  37.2727273 208.9090909 124.0909091
107.8181818 133.4545455
```

(ii) a new data frame containing just individuals with bmp in the interval [70,90]

```
> b = a[bmp>70 & bmp<90, ]
> apply(b,2, mean)
        age         sex      height      weight
bmp        fev1          rv         frc         tlc
pemax
 17.5000000   0.3333333 163.0000000  46.2500000
82.1666667  32.0000000 226.6666667 139.5000000
107.3333333 100.0000000
```

(iii) a new data frame containing just individuals with fev1 > 30 or rv > 300

```
> b = a[fev1>30 | rv>300, ]
> apply(b,2, mean)
        age         sex      height      weight
bmp        fev1          rv         frc         tlc
pemax
 14.0476190   0.3333333 151.5238095  37.4238095
77.7619048  36.3333333 262.8095238 158.4761905
114.1904762 109.4761905
```

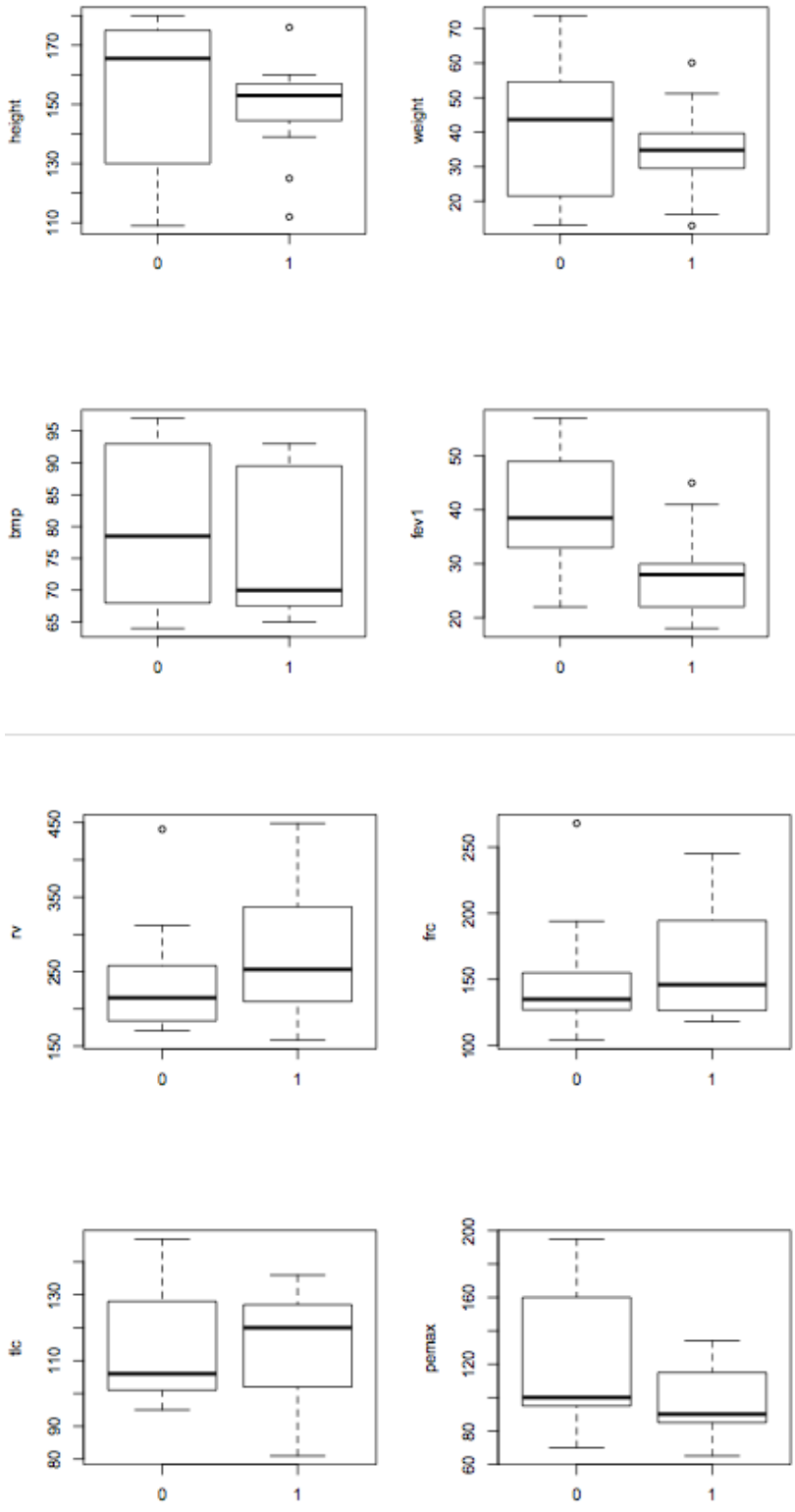
Q6. Plot a histogram for the variable height and overlay the density estimate on to the histogram using a blue line (hint : blue is col = 4)

```
> hist(height, freq = FALSE)
> lines(density(height), col = 4)
```

Q7. Plot histograms for height for each sex separately, one above the other. Make sure the x-axis has the same range on both plots.

```
par(mfrow = c(2,1))
hist(height[sex==0], xlim = c(100, 200))
hist(height[sex==1], xlim = c(100, 200))
```

Q8. Create boxplots for the variables height, weight, bmp, fev1, rv, frc, tlc and pemax, all stratified by sex. Which have evidence of outlying observations



Q9. Use scatterplots between the variables to find any clear relationships between the variables?

Juul dataset

Q10. Read in the data from the file `juul.txt` as a data frame and attach to it.

```
a = data.frame(read.table(file ="juul.txt",hea=T))
attach(a)
```

Q11. Create summaries of the variables in this dataset?

```
summary(a)
```

Q12. Which variable has the most missing data?

```
testvol
```

Q13. How many individuals of each sex are there in the dataset?

```
> sum(sex==1, na.rm = T)
[1] 621
> sum(sex==2, na.rm = T)
[1] 713
```

Q14. Use the `table` command to create a contingency table of the factors sex and tanner.

```
t1=table(sex,tanner)
```

Q15. Produce a barplot with a bar for each level of the factor tanner where each bar is split into the two levels of the factor sex.

```
barplot(t1, legend = TRUE)
```

Q16. Produce another barplot where the roles of the variables sex and tanner are reversed.

```
barplot(t(t1), legend = TRUE,ylim = c(0, 800))
```

Simulation examples

Q17. (plotting) The `rgamma(n,a,b)` function simulates `n` Gamma`(a,b)` rv. Simulate 10000 Gamma `(3,4.2)` rv and make a histogram. Overlay a plot of the density. (`dgamma(x,a,b)` is the Gamma(a,b) density)

```
X<-rgamma(10000,3,4.2)
a<-hist(X,freq=F,100)
```

```
lines(a$breaks,dgamma(a$breaks,3,4.2))
```

Q18. (vectorising) Simulate a Gaussian random walk with 100 steps: `X[1]=0, X[i]=X[i-1]+rnorm(1)`. Plot the walk. Try to vectorise your code.

```
X<-cumsum(c(0,rnorm(99)))
plot(X,type='l')
```

Q19. What does the following code do? What happened to the `while()` loop?

```
n<-1000;
y<- log(runif(n))* sample(c(-1,1),n,replace=T);
u<-runif(n);
p.over.Mq<-exp(-y^2/2+abs(y)-0.5)
X=y[u<p.over.Mq]
qqnorm(X); qqline(X)
```

This is a form of rejection – as it was first explained. We sample points `(y,uMq(y))` uniformly at random under the curve of `Mq(y)` and accept them if they lie under `p(y)`.