MSIT 431: PROBABILITY AND STATISTICAL METHODS

Submitted By : Ruma Anand

Student ID : 3074142

EXERCISES - SECTION 2

1. Define

> x<-c(4,2,6)

> y<-c(1,0,-1)

Decide what the result will be of the following:

(a)> length(x)

Ans. [1] 3

(b)> sum(x)

Ans. [1] 12

(c) > sum(x^2)

Ans. [1] 56

(d) > x+y

Ans. [1] 5 2 5

(e) > x*y

Ans. [1] 4 0 -6

(f) > x-2

Ans. [1] 2 0 4

(g) > x^2

Ans. [1] 16 4 36

2. Decide what the following sequences are and use R to check your answers:

(a)> 7:11

Ans. [1] 7 8 9 10 11

(b) > seq(2,9)

Ans. [1] 2 3 4 5 6 7 8 9

(c) > seq(4,10,by=2)

Ans. [1] 4 6 8 10

(d) > seq(3,30,length=10)

Ans. [1] 3 6 9 12 15 18 21 24 27 30

(e) > seq(6,-4,by=-2)

Ans. [1] 6 4 2 0 -2 -4

3. Determine what the result will be of the following R expressions, and then use R to check you are right:

(a) > rep(2,4)

Ans. [1] 2 2 2 2

(b) > rep(c(1,2),4)

Ans. [1] 1 2 1 2 1 2 1 2

(c) > rep(c(1,2),c(4,4))

Ans. [1] 1 1 1 1 2 2 2 2

(d) > rep(1:4,4)

Ans. [1] 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4

(e) > rep(1:4,rep(3,4))

Ans. [1] 1 1 1 2 2 2 3 3 3 4 4 4

4. Use the rep function to define simply the following vectors in R.

(a) 6 6 6 6 6 6

Ans. > rep(6,6)

(b) 5 8 5 8 5 8 5 8

> rep(c(5,8),4)

(c) 5 5 5 5 8 8 8 8

Ans. > c(rep(5,4),rep(8,4))

EXERCISES - SECTION 3

5. If x<- c(5,9,2,3,4,6,7,0,8,12,2,9) decide what each of the following is and use R to check your answers:

(a) > x[2]

Ans. [1] 9

(b) > x[2:4]

Ans. [1] 9 2 3

(c) > x[c(2,3,6)]

Ans. [1] 9 2 6

(d) > x[c(1:5,10:12)]

Ans. 5 9 2 3 4 12 2 9

(e) > x[-(10:12)]

Ans. [1] 5 9 2 3 4 6 7 0 8

2. The data y<-c(33,44,29,16,25,45,33,19,54,22,21,49,11,24,56) contain sales of milk in litres for 5 days in three different shops (the first 3 values are for shops 1,2 and 3 on Monday, etc.) Produce a statistical summary of the sales for each day of the week and also for each shop.

Ans.

```
> Monday<-y[1:3]
> summary(Monday)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  29.00   31.00   33.00   35.33   38.50   44.00
> Tuesday<-y[4:6]
> summary(Tuesday)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  16.00   20.50   25.00   28.67   35.00   45.00
> Wednesday<-y[7:9]
> summary(Wednesday)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  19.00   26.00   33.00   35.33   43.50   54.00
> Thursday<-y[10:12]
> summary(Thursday)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  21.00   21.50   22.00   30.67   35.50   49.00
> Friday<-y[13:15]
> summary(Friday)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  11.00   17.50   24.00   30.33   40.00   56.00
> Shop1<-(y[c(1,4,7,10,13)])
> summary(Shop1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     11      16      22      23      33      33
> Shop2<-y[c(2,5,8,11,14)]
> summary(Shop2)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   19.0    21.0    24.0    26.6    25.0    44.0
> Shop3<-y[c(3,6,9,12,15)]
> summary(Shop3)
```

Min. 1st Qu. Median Mean 3rd Qu. Max.

29.0 45.0 49.0 46.6 54.0 56.0

EXERCISES - SECTION 5

1. Attach to the dataset quakes and produce a statistical summary of the variables depth and mag.

Ans.

> attach(quakes)

> summary(depth)

Min. 1st Qu. Median Mean 3rd Qu. Max.

40.0 99.0 247.0 311.4 543.0 680.0

> summary(mag)

Min. 1st Qu. Median Mean 3rd Qu. Max.

4.00 4.30 4.60 4.62 4.90 6.40

2. Attach to the dataset mtcars and find the mean weight and mean fuel consumption for vehicles in the dataset (type help(mtcars) for a description of the variables available).

Ans.

> attach(mtcars)

> help(mtcars)

starting httpd help server ... done

> mean(mpg)

[1] 20.09062

> mean(wt)

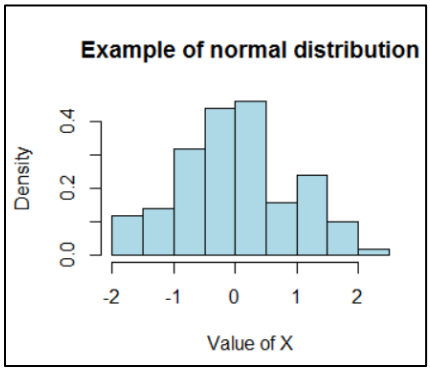
[1] 3.21725

EXERCISES - SECTION 8

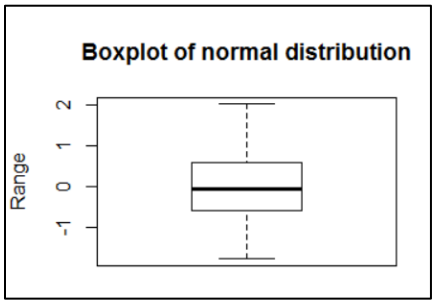
1. Use > x<-rnorm(100) or something similar, to generate some data. Produce a figure showing a histogram and boxplot of the data. Modify the axis names and title of the plot in an appropriate way.

Ans.

hist(x,freq=FALSE,xlab="Value of X", main="Example of normal distribution",col="lightblue")



boxplot(x,main="Boxplot of normal distribution",ylab="Range")



2. Type the following:

> x<- (-10):10

> x

[1] -10 -9 -8 -7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8

[20] 9 10

> n<-length(x)

> n

[1] 21

> y<-rnorm(n,x,4)

> y

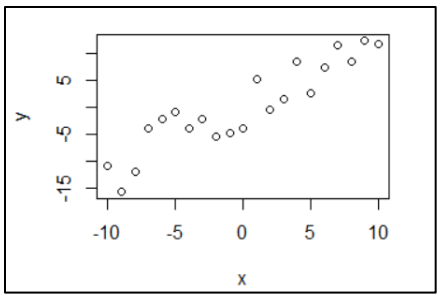
[1] -10.7522596 -15.6920998 -11.9417310 -3.8033309 -2.1768232 -0.9494487

[7] -3.8391703 -2.0804432 -5.3793581 -4.8516299 -3.8645589 5.1752916

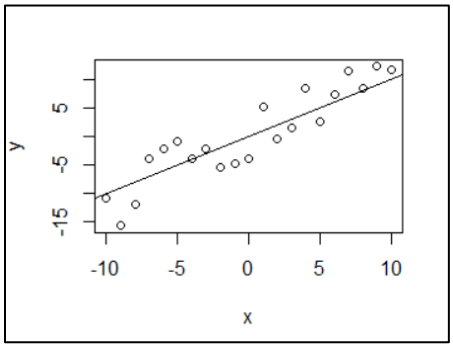
[13] -0.3499266 1.5534286 8.4624597 2.5727006 7.4098451 11.4801652

[19] 8.4175258 12.2780979 11.7355825

> plot(x,y)



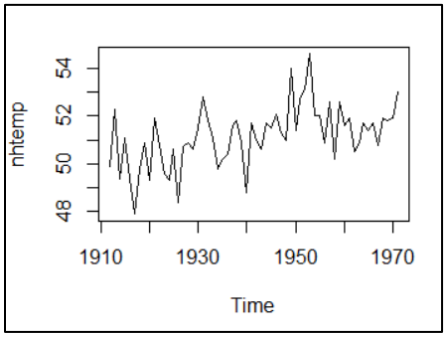
> abline(0,1)



3. Type the following :-

> data(nhtemp)

> plot(nhtemp)



---

5. Consider the following data set:

(97 82 85 61 14 22 51 16 55 44 70)

Given the 5 number summary (as there are slightly different ways to define Q1 and Q3, you need to clarify the definition you use if you don't follow strictly the textbook's definition);

Is there any outlier?

Give the mean and standard deviation of the data.

Ans.

Let x<-c(97,82,85,61,14,22,51,16,55,44,70)

> mean(x)

[1] **54.27273**

> sd(x)

[1] **28.41862**

summary(x)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
14.00	33.00	55.00	54.27	76.00	97.00

The interquartile range IQR =Q3-Q1 = 43 and 43*1.5=64.5(IQR rule for outliers).

As per the 1.5* IQR rule for outliers, any values below -31.5(33-64.5) and above 140.5(76+64.5) should be an outlier. Since there are no values in this range, **there are no outliers**.

---

6. Textbook Exercise 1.128

Find an SAT percentile. Reports on a student's ACT and SAT results usually give the percentile as well as the actual score. The percentile is just the cumulative proportion stated as a percent: the percent of all scores that were lower than or equal to this one. Renee scores 2050 on SAT. What is her percentile?

Ans

As informed earlier, the distribution of SAT scores is approximately Normal with mean μ=1498 and standard deviation σ=316.

Hence,

$$z=(x-\mu)/\sigma$$

$$z=(2050-1498)/316$$

$$z=1.75$$

Standard normal probability table yield z=1.75 is .96 or 96th percentile.

---

6. Textbook Exercise 1.162

A university department installed a spam filter on its computer system. During a 21-day period, 6693 messages were tagged as spam. How much spamyou get depends on what your online habits are. Here are the counts for some students and faculty in this department. All other department members received fewer than 100 spam messages. How many did the others receive in total. Make a graph and comment on what you learn from these data.

Ans.

```
> z=data.frame(ID=c('AA','BB','CC','DD','EE','FF','GG','HH','II','JJ','KK','LL'),
+ Count = c(1818,1358,442,416,399,389,304,251,251,178,158,103))
> z
   ID Count
1  AA  1818
```

2 BB 1358

3 CC 442

4 DD 416

5 EE 399

6 FF 389

7 GG 304

8 HH 251

9 II 251

10 JJ 178

11 KK 158

12 LL 103

> sum(z[,2])

[1] 6067

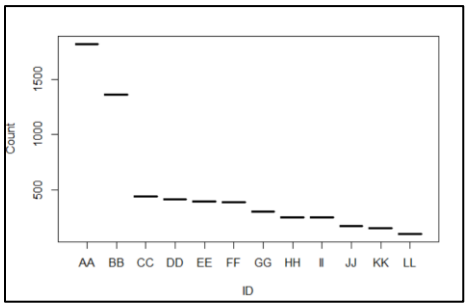
Total messages : 6693

Messages received by other departments

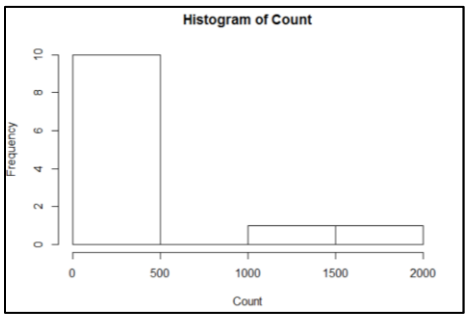
> 6693-6067

[1] 626

> plot(z)



> hist(Count)



We learn from the graph and histogram that out of 6693 spam messages received, student/faculty with ID as AA and BB had exceedingly high count of spam messages. They may be outliers but we cannot say for sure since we do not have the data for remaining spam messages. Also, the frequency of spam messages for members other than AA and AB were all under 500.