R Language

About R Language

- R is a computer language for carrying out statistical computations.
- R is Free Software, and runs on a variety of platforms
- Command-line execution based on function calls.
- Workspace containing data and functions.
- Extensible with user functions.
- Graphics devices.
- R packages can contain not only code, but also other resources like documentation and sample data sets.
- Well-defined format that ensures easy installation, a basic standard of documentation, and enhances portability and reliability.

About R Language

- The basic mode of interaction is ‘read – evaluate – print’.
- ▫ The R Project is an international collaboration of researchers in statistical computing.
- ▫ There are roughly 20 members of the “R Core Team” who maintain and enhance R.
- ▫ Releases of the R environment are made through the CRAN (comprehensive R archive network) twice per year.
- ▫ The software is released under a “free software” license, which makes it possible for anyone to download and use it.
- ▫ R is a computer language which is processed by a special program called an interpreter. This program reads and evaluates R language expressions, and prints the values determined for the expressions.
- ▫ There are over 3500 extension packages that have been contributed to CRAN.

Installation

- ▫ R can be downloaded from one of the mirror sites in http://cran.r- project.org/mirrors.html.

Using External Data

- ▫ R offers plenty of options for loading external data, including Excel, Minitab, SAS and SPSS files.

R Date and Time Functions

- R has serveral date and time related functions. **date()** functions returns a date without time as character string. **Sys.Date()** and **Sys.time()** returns the system's date .

```
>date()
[1] "Fri Jan 04 17:38:05 2017"
>Sys.time()
[1] "2017-01-04 17:47:39 IST"
>Sys.Date()
[1] "2017-01-04"
>class(date())
 [1] "character"
>class(Sys.Date())
 [1] "Date"
```

sequence of numbers

- The expression n1:n2, generates the sequence of integers from n1 to n2.

```
> 1:15    #print the numbers 1 to 15
 [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15
```

```
> 5:-5# print the numbers 5 to -5
 [1]  5  4  3  2  1  0 -1 -2 -3 -4 -5
```

seq() function generates a sequence of numbers.

- Generate a sequence from -6 to 7:

```
 > x <- seq(-6,7)
> x
[1] -6 -5 -4 -3 -2 -1  0  1  2  3  4  5  6  7
```

sequence of numbers

- From -6 till 7, step=2:

```
> x <- seq(-6,7,by=2)
> x
[1] -6 -4 -2  0  2  4  6
```

sequence of numbers

```
> x <- seq(-2,2,by=0.3)
> x
[1] -2.0 -1.7 -1.4 -1.1 -0.8 -0.5 -0.2  0.1  0.4 0.7  1.0  1.3  1.6  1.9
```

Suppose we do not know the step, but we want 10 evenly distributed numbers from -2 to 2:

```
> seq(-2,2,length.out=10)
[1] -2.0000000 -1.5555556 -1.1111111 -0.6666667 -0.2222222  0.2222222
[7]  0.6666667  1.1111111  1.5555556  2.0000000
```

Generate a sequence from 1 to 10, quick version:

```
> x <- seq(10)
> x
[1]  1  2  3  4  5  6  7  8  9 10
```

exp(x) function compute the exponential value of a number or number vector, ex.

```
> x <- 5
> exp(x)
[1] 148.4132
```

R rep Function

rep() function replicates the values in x.

```
>x <- rep(1:5)
[1] 1 2 3 4 5
```

Repeat 1 -5 two times:

```
>x <- rep(1:5,2)
[1] 1 2 3 4 5 1 2 3 4 5
```

```
> x=rep(1:10)
> exp(x)
 [1]     2.718282     7.389056    20.085537    54.598150   148.413159
 [6]   403.428793  1096.633158  2980.957987  8103.083928 22026.465795
```

Data Frame:

- A **data frame** is used for storing data tables. It is a list of vectors of equal length

```
> patientID <- c(1, 2, 3, 4)
> age <- c(25, 34, 28, 52)
> diabetes <- c("Type1", "Type2", "Type1", "Type1")
> status <- c("Poor", "Improved", "Excellent", "Poor")
> patientdata <- data.frame(patientID, age, diabetes, status)
> patientdata
  patientID age diabetes    status
1         1  25    Type1      Poor
2         2  34    Type2  Improved
3         3  28    Type1 Excellent
4         4  52    Type1      Poor
`
```

Listing 2 Specifying columns of a data frame

```
> patientdata[1:2]
  patientID age
1         1  25
2         2  34
3         3  28
4         4  52
> patientdata[c("diabetes", "status")]
  diabetes    status
1    Type1      Poor
2    Type2  Improved
3    Type1 Excellent
4    Type1      Poor
> patientdata$age    #age variable in the patient data frame
[1] 25 34 28 52
```

Arithmetic Mean (or Simply Mean):

It is defined as the sum of the given observations divided by the total number of observations.

$$\text{Arithmetic Mean (A.M.)} = \bar{X} = \frac{\text{Sum of all observations}}{\text{Total number of observations}}$$

$$\bar{X} = \frac{\sum X}{n}$$

where $\sum X$ = Sum of all observations. (Read $\sum$ as capital Sigma)

n = Total number of observations.

Case A : Raw Data

Let $X_1, X_2, \ldots, X_n$ be 'n' measurements. The arithmetic mean of this data set can be computed by using formula:

$\bar{X} = \frac{\sum X}{n}$, where $\sum X = X_1 + X_2 + \ldots, + X_n$.

n = No. of observations in the given data.

Case B: Discrete frequency distribution

Consider the following discrete frequency distribution of variable values and their corresponding frequencies

Variable Value *(X)*	X_1	X_2	...	X_k	*Total*
Frequency *(f)*	f_1	f_2	...	f_k	N

The Arithmetic mean is then defined as,

$\bar{X} = \frac{\sum fX}{N}$, where $\sum fX = f_1X_1 + f_2X_2 + .. + f_kX_k$

N = Total Frequency ($\sum f$)

Case C: Continuous frequency distribution

In this case, A.M. is given by $\bar{X} = \frac{\sum fX}{N}$,

where $\sum fX$ = Sum of products of midvalues of class intervals and the corresponding frequencies.

N = Total frequency. When mid values of class intervals are large in magnitude, the step deviation method (or short cut method) can be employed to find A.M.

Problem1: Twenty students , graduates and undergraduates, were enrolled in a statistics course. Their ages were

18,19,19,19,19,20,20,20,20,20,21,21,21,21,22,23,24,27,30,36.

a) *Find Mean and Median of all students*
b) *Find median age of all students under 25 years.*
c) *Find modal age of all students*

R code:-

```
> x=c(18,19,19,19,19,20,20,20,20,20,21,21,21,21,22,23,24,27,30,36)
> mean(x)   #mean
[1] 22
> median(x)   #median
[1] 20.5
> y=x[x<25]
> md=median(y)
> md
[1] 20
> xr=table(x)    #mode
> mode=which(xr==max(xr))
> mode
20
```

Problem 2 : A survey of 25 faculty members is taken in a college to study their vocational mobility.They were asked the question "In addition to your present position ,at how many educational instistutes have srved on the faculty?.Following is the frequency distribution of their responses.

X	0	1	2	3
f	8	11	5	1

Find mean and median of the distribution

R code:

```
> x=c(0,1,2,3)
> f=c(8,11,5,1)
> y=rep(x,f)
> mean=(sum(y))/(length(y))    #mean
> mean
[1] 0.96
> median(y)       #median
[1] 1
```

Problem 3 : Compute mean ,median and mode of for the following frequency

Distribution:

Height in Cm	*145-150*	*150-155*	*155-160*	*160-165*	*165-170*	*170-175*	*175-180*	*180-185*
No. of Adult men	*4*	*6*	*28*	*58*	*64*	*30*	*5*	*5*

R code:-

```
> mid=seq(147.5,182.5,5)
> mid
[1] 147.5 152.5 157.5 162.5 167.5 172.5 177.5 182.5
> f=c(4,6,28,58,64,30,5,5)
> fr.distr=data.frame(mid,f)
> fr.distr
   mid  f
1 147.5  4
2 152.5  6
3 157.5 28
4 162.5 58
5 167.5 64
6 172.5 30
7 177.5  5
8 182.5  5
```

Mean:--

```
> mean=(sum(mid*f))/sum(f)
> mean
[1] 165.175
Median

> midx=seq(147.5,182.5,5)
> frequency=c(4,6,28,58,64,30,5,5)
> fr.dist<-data.frame(midx,frequency)
> fr.dist
   midx frequency
1 147.5         4
2 152.5         6
```

```
3 157.5       28
4 162.5       58
5 167.5       64
6 172.5       30
7 177.5        5
8 182.5        5
> cl=cumsum(frequency)
> cl
[1]   4  10  38  96 160 190 195 200
> n=sum(frequency)
> n
[1] 200
> ml=min(which(cl>=n/2))        # The serial number of the median class
> ml
[1] 5
> h=5
> h
[1] 5
> f=frequency[ml]                    #frequency of the median class
> f
[1] 64
> c=cl[ml-1]
> c
[1] 96
> l=mid[ml]-h/2
> l
[1] 165
> median=l+(((n/2)-c)/f)*h                    #median
> median
[1] 165.3125
```

```
Mode:-
> m=which(frequency==max(frequency)) #serial number of the median class
> m
[1] 5
> fm=frequency[m]     # frequency of the modal class
> fm
[1] 64
> f1=frequency[m-1]  # frequency of the pre modal class
> f2=frequency[m+1] # frequency of the post modal class
> f1
[1] 58
> f2
[1] 30
> l=midx[m]-h/2
> l
[1] 165
> mode=l+((fm-f1)/(2*fm-f1-f2))*h
> mode
[1] 165.75
```

Digital Assignment 1

Take a survey of your class students with X as a number of whats app messages that respond by the students during any one of the following:

(i) During statistics period

(ii) During 5 am to 6am

(iii) During 1pm to 2pm

(iv) During 10 pm to 11pm

Measure of dispersion :-

The various measures of absolute variation are (i) Range (ii) Quartile Deviation (iii) Mean Deviation and (iv) Standard Deviation.

Range = Largest value – Smallest value

QUARTILE DEVIATION (Q.D):

$$Q.D = \frac{Q_3 - Q_1}{2}$$

'Coefficient of Quartile Deviation'

$$Q.D = \left[\frac{Q_3 - Q_1}{Q_3 + Q_1}\right]$$

Mean deviation about 'A' is

$$\text{M.D.} = \frac{\sum|x-A|}{n} \text{ where } |x\text{-}A| \text{ is an absolute deviation taken from A}$$

STANDARD DEVIATION (S.D):

$$\text{Direct formula : } \sigma = \sqrt{\frac{\sum(x-\bar{x})^2}{n}} \text{ where } \bar{x} = \frac{\sum x}{n} = A.M.$$

An entomologist studying morphological variation in species of mosquito recorded the following data on body length:

1.2,1.4,1.3,1.6,1.0,1.5,1.7,1.1,1.2,1.3

Compute all the measures of disersion.

```
> x=c(1.2,1.4,1.3,1.6,1.0,1.5,1.7,1.1,1.2,1.3)
> x
 [1] 1.2 1.4 1.3 1.6 1.0 1.5 1.7 1.1 1.2 1.3
> summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.200   1.300   1.330   1.475   1.700
> range=1.7-1.0                #range
> range
[1] 0.7
> var(x)                       #variance
[1] 0.049
> sd=sqrt(var(x))              #standard deviation
> sd
[1] 0.2213594
```

r^{th} Population Moment about Mean=$\mu_r = \dfrac{\sum (x_i - \mu)^r}{N}$

the coefficient of skewness is given by $\beta_1 = \dfrac{\mu_3^2}{\mu_2^3}$

Beta coefficient of kurtosis (β_2) $\beta_2 = \dfrac{\mu_4}{\mu_2^2}$

Fisher's Gamma coefficient of kurtosis (γ_2) $\gamma_2 = \beta_2 - 3$

Problem : A quality control engineer is intrested in determining whether a machine is properly adjusted to dispense 16 ounces of sugar.Following data refer to the net weight(in ounces) packed in thirty one-pound bags after the machine was adjusted.Compute the measures skewness and kurtosis

15.9,16.2,16.0,15.6,16.2,15.9,16.0,15.6,15.6,16.0,16.2,15.6,15.9,16.2,15.6,16.2,
15.8,16.0,15.8,15.9,16.2,15.8,15.8,16.2,16.0,15.9,16.2,16.2,16.0,15.6

```
>x=c(15.9,16.2,16.0,15.6,16.2,15.9,16.0,15.6,15.6,16.0,16.2,15.6,15.9,16.2,15.6,16.2,15.8,16
.0,15.8,15.9,16.2,15.8,15.8,16.2,16.0,15.9,16.2,16.2,16.0,15.6)
> x
 [1] 15.9 16.2 16.0 15.6 16.2 15.9 16.0 15.6 15.6 16.0 16.2 15.6 15.9 16.2 15.6
[16] 16.2 15.8 16.0 15.8 15.9 16.2 15.8 15.8 16.2 16.0 15.9 16.2 16.2 16.0 15.6
> n=length(x)
> n
[1] 30
> mean=mean(x)
> mean
[1] 15.93667
```

```
> m4=sum((x-mean)^4)/n
> m4
[1] 0.004062022
> m2=var(x)
> m2
[1] 0.0486092
> beta2=m4/(m2^2)
> beta2
[1] 1.719117
> gam2=beta2-3
> gam2
[1] -1.280883
```

ASSESSMENT I

Upload your Assessment I, the screen shot of the workspace on or before the due date(refer vtop)

- Exercise1:

  Calculate the mean, median and mode for the following data:

Class	130-135	135-140	140-145	145-150	150-155	155-160	160-165
Frequency	5	15	28	24	17	10	1

- Exercise 2:

  Calculate the mean, median and mode

Size	0-4	4-8	8-12	12-16	16-20
Frequency	10	20	30	25	20

ASSESSMENT-I (cont…)

- Exercise 3:

  Find the mean, standard deviation and quartile deviation of the following data

Class	15-25	25-35	35-45	45-55	55-65	65-75
Frequency	7	15	20	18	14	6