LAB-3

Descriptive Statistics

(Measure of central tendency, Measure of dispersion, Measure of Skewness and Measure of kurtosis)

Description:-

MEASURES OF CENTRAL TENDENCY

A measure of central tendency is a typical value that serves as a representative of all the measurements. The measurements obtained from a common source are not likely repetitions. It is undesirable to keep all the measurements in focus. Consequently, a representative of all the measurements is required. Such a representative is one of the three popular measures of central tendency. Viz., arithmetic mean, median and mode. A measure of central tendency is a numerical value around which the measurements have a tendency to cluster.We come across five measures of central tendency (i) arithmetic mean, (ii) median, (iii) mode, (iv) geometric mean and (v) harmonic mean.

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 $\sum fX = f_1X_1 + f_2X_2 + \ldots + f_kX_k$

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]    #mode
> md=median(y)
> md
[1] 20
> xr=table(x)    #mode
> mode=which(xr==max(xr))
> mode
20
 3
```

Measures of central tendency for frequency table:-

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]                     # cumulative frequency of the median class
> 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
```

Measure of dispersion :-

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

RANGE :

The simplest but a crude measure of variation is range. it is defined as the difference between the highest and lowest values in the series (data). Suppose a raw data set contains 'n' observations. Then the Range is the difference between the largest and smallest values in the data.

Range = Largest value – Smallest value

QUARTILE DEVIATION (Q.D):

It is a better measure of variation than range. It is based on the first and third quartiles, namely Q_1 and Q_3 It is defined as,

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

The computation of quartile deviation is based on the computations of Q_3 and Q_1.

It is superior and more reliable measure than the range as it makes use of 50% of the data. When a frequency distribution contains open end classes, Q.D. is the appropriate measure of dispersion. Q.D. is a measure of absolute dispersion. A measure of relative dispersion based on Q.D. is known as "Coefficient of Quartile Deviation" which is defined as coefficient of $Q.D = \left[\frac{Q_3 - Q_1}{Q_3 + Q_1}\right]$. It is useful to compare the variations in two or more series of data.

Mean Deviation (M.D.)

It is defined as the arithmetic mean of absolute deviations (obtained by ignoring the sign) of observations taken from an average (Mean, Median or Mode). Thus we have

(a) mean deviation about arithmetic mean

(b) mean deviation about median and

(c) mean deviation about mode.

Some times we may choose any arbitrary value 'A' in the place of an average.

Suppose a data set contains n observations say $x_1, x_2, \ldots\ldots x_n$. Let A be any arbitrary value (A may be A.M. or Median or Mode or any arbitrary value), then Mean deviation about 'A' is defined, by

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

STANDARD DEVIATION (S.D):

It is defined as the positive square root of the arithmetic mean of the squares of the deviation of the observations taken from their arithmetic mean. It is usually denoted by the Greek letter (small sigma) σ. The square of standard deviation is known as the variance of data (σ^2). Suppose raw data contain 'n' observations, say, $x_1, x_2, \ldots, x_n$. The standard deviation is defined as .

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

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.

R code:-

```
> 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

There is  no separate command for Quartile deviation Mean deviation .We
have to evaluate the expression

>cqd=(1.475-1.2)/(1.475+1.2)       #coefficient of quartile deviation

# Mean deviation about Mean
> y=(x-mean(x))
> y
 [1] -0.13  0.07 -0.03  0.27 -0.33  0.17  0.37 -0.23 -0.13 -0.03
> y=abs(y)
> y
 [1] 0.13 0.07 0.03 0.27 0.33 0.17 0.37 0.23 0.13 0.03
> mdl=sum(y)/length(y)
> mdl
[1] 0.176

#Mean deviation about Median

> z =abs(x-median(x))
> md2=sum(z)/length(z)
> md2
```

[1] 0.17

Mean deviation about Mode

 # in this Problem ,it is a bi-model series (Mode is not possible)

Measure of skewness and kurtosis using Moments:

Moments measure of skewness: It is a measure based on central moments. By this method, the coefficient of skewness is given by

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}, \gamma_1 = \sqrt{\beta_1}$$

$\beta_1 = 0 \Rightarrow$ symmetric distribution

$\beta_1 > 0$ positively skewed distribution

$\beta_1 < 0$ negatively skewed distribution

Beta coefficient of kurtosis (β_2)

Karl pearson suggested a measure of kurtosis (β_2), which is based on second and fourth central moments. $\beta_2 = \frac{\mu_4}{\mu_2^2}$

where μ_2 and μ_4 are the central moments of order two and four respectively

$\beta_2 = 3 \Rightarrow$ Curve is a measokuritic curve or normal curve

$\beta_2 = 3 \Rightarrow$ Curve is a Leptokurtic curve

$\beta_2 = 3 \Rightarrow$ Curve is a platykurtic curve

(ii) Fisher's Gamma coefficient of kurtosis (γ_2)

Fisher introduced γ_2 coefficient as a measure of kurtosis.It is given by $\gamma_2 = \beta_2 - 3$

$\gamma_2 = 0 \Rightarrow$ Mesokurtic curve or normal curve

$\gamma_2 > 0 \Rightarrow$ Leptokurtic curve

$\gamma_2 < 0 \Rightarrow$ platykurtic curve

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

R code:-

```
>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
```

Experiment:-

Collect at least 60 students and analyse the data by using descriptive statistics and Interpret the results.