Lab-4

CORRELATION

Correlation Definition :-

Correlation refers to the relationship between two or more variables. Simple correlation studies the relationship between two variables. Correlation analysis attempts to determine the degree of relationship between variables.

Measures of Correlation:

Scatter Diagram:

Scatter diagram is the simplest way of graphic representation of a bivariate data, where the given set of 'n' pairs of observations on two variables X and Y say (X_1, Y_2), (X_2, Y_2) …(X_n, Y_n) may be plotted as dots by considering X-values on X-axis and Y-values on Y-axis. By scatter diagram, we can get some idea about the correlation between X and Y.

Problem:-

AGE GROUP	REPRESENTATIVE AGE	HOURS SPEND IN THE LOCAL LIBRARY
10-19	15	302.38
20-29	25	193.63
30-39	35	185.46
40-49	45	198.49
50-59	55	224.30
60-69	65	288.71

illustrate the relationship between the average age versus the time spent in the library, by using scatterplot.

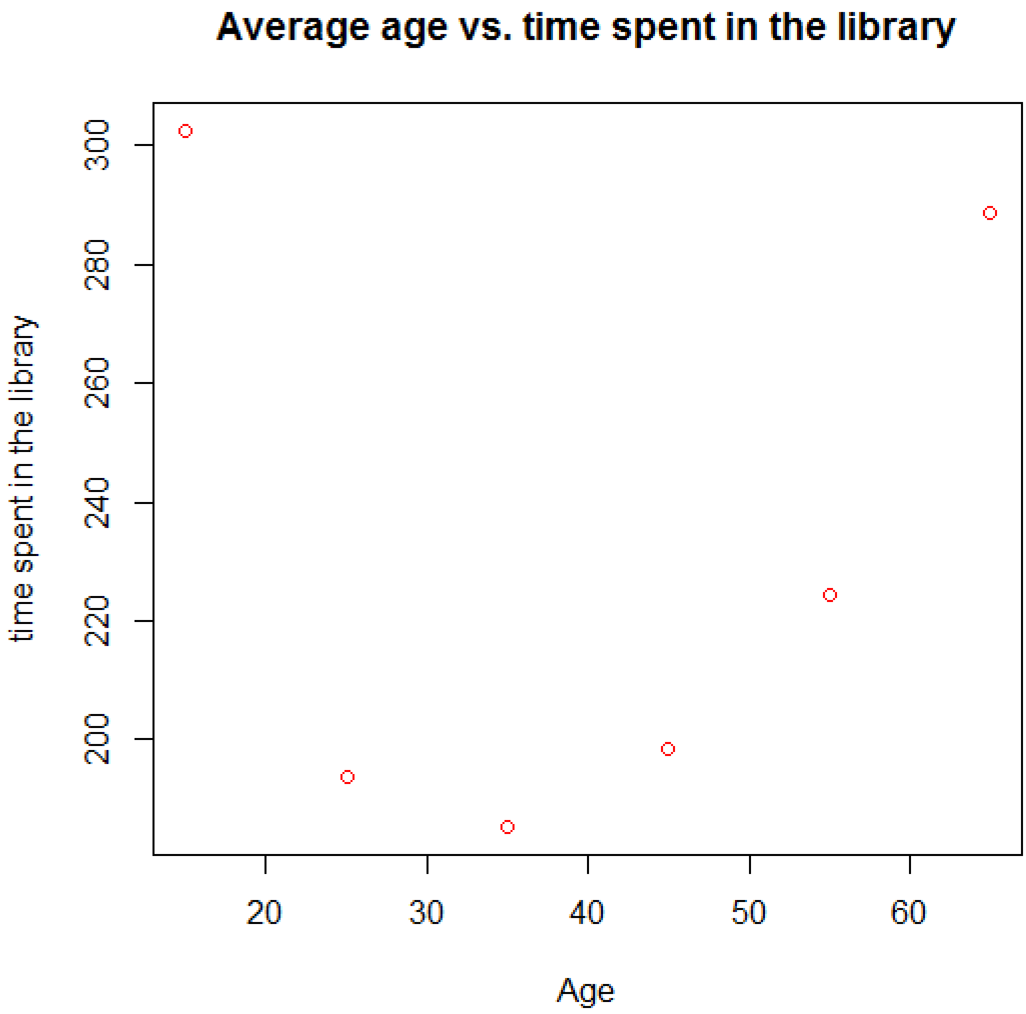
R code:-

```
>x <- c(15,25,35,45,55,65)

> y <- c(302.38, 193.63, 185.46, 198.49, 224.30, 288.71)
```

> plot(x,y, main="Average age vs. time spent in the library", xlab="Age", ylab="time spent in the library",col="red")

OUTPUT:-



Karl Pearson's Coefficient of Correlation

It is defined as the ratio of covariance between x and y say Cov (X,Y) to the product of the standard deviations of X and Y, say $\sigma_X \sigma_Y$

$$i.e \qquad r_{XY} = \frac{Cov(XY)}{\sigma_X \sigma_Y}$$

Consider a set of 'n' pairs of observations $(X_1, Y_1), (X_2, Y_2), \ldots (X_n, Y_n)$ *on two variables X and Y. Then we have, Covariance between X and Y*

R code:-

```
> x=c(23,27,28,28,29,30,31,33,35,36)
> y=c(18,20,22,27,21,29,27,29,28,29)
> var(x)
[1] 15.33333
> var(y)
[1] 18.22222
> var(x,y)
[1] 13.66667
> r=var(x,y)/sqrt(var(x)*var(y))
> r
[1] 0.8176052
Or
> cor(x,y)
[1] 0.8176052
Or
> cor.test(x,y) Or
> cor.test(x,y,method="pearson")
        Pearson's product-moment correlation
data:  x and y
t = 4.0164, df = 8, p-value = 0.003861
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3874142 0.9554034
sample estimates:
      cor
0.8176052
```

There is a Positive correlation between X and Y

SPEARMAN'S RANK CORRELATION COEFFICIENT

Suppose we associate the ranks to individuals or items in two series based on order of merit, the Spearman's Rank correlation coefficient ρ is given by

$$\rho = 1 - \left[\frac{6\sum d^2}{n(n^2-1)}\right]$$ [Read the symbol (as 'Rho'.]

Where, $\sum d^2$ = Sum of squares of differences of ranks between paired items in two series n = Number of paired items'

SPEARMAN'S RANK CORRELATION COFFICIENT FOR A DATA WITH AND WITHOUT TIED OBSERVATIONS:

Problem : Twelve recruits were subjected to selection test to ascertain their suitability for a certain course of training. At the end of training they were given a proficiency test. The marks scored by the recruits are recorded below :

Recruit	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
Selection Test Score	*44*	*49*	*52*	*54*	*47*	*76*	*65*	*60*	*63*	*58*	*50*	*67*
Proficiency Test Scrore	*48*	*55*	*45*	*60*	*43*	*80*	*58*	*50*	*77*	*46*	*47*	*65*

Calculate rank correlation coefficient and comment on your result.

Solution:-

```
> selection =c(44,49,52,54,47,76,65,60,63,58,50,67)
> proficiency =c(48,55,45,60,43,80,58,50,77,46,47,65)
> cor.test(selection,proficiency,method ="spearman")

        Spearman's rank correlation rho

data:  selection and proficielncy
S = 80, p-value = 0.01102
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.7202797
```

There is a positive correlation between selection and Proficiency

KENDALL'S COEFFICIENT OF CONCURRENT DEVIATIONS

The Kendall's coefficient of concurrent deviations is denoted by r_c and defined as

$$r_c = \pm\sqrt{\pm\left[\frac{2C - n}{n}\right]}$$

Where, C = Number of concurrent deviations or position signs of (D_X, D_Y);

n = Number of pairs of deviations

Problem: The following data gives the marks obtained by 12 students in statistics and computer science :

Students		1	2	3	4	5	6	7	8	9	10	11	12
Marks	Statistics	55	40	70	60	62	73	65	65	20	35	46	50
	Computer Science	35	32	65	50	63	45	50	65	70	72	72	40

Compute the coefficient of correlation by the method of concurrent deviations.
R code:

```
> statistics=c(55,40,70,60,62,73,65,65,20,35,46,50)

> mathematics=c(35,32,65,50,63,45,50,65,70,72,72,40)

> cor.test(statistics,mathematics,method="kendall")

Kendall's rank correlation tau

data:  statistics and mathematics

z = -0.27688, p-value = 0.7819

alternative hypothesis: true tau is not equal to 0

sample estimates:

      tau

-0.06250763

There is a negative correlation between mathematics and statistics
```

R^2 *(Coefficient of determination):-*

Code:

examdata=read.csv("C:\\Users\\aadmin\\Desktop\\mokesh\\examdata.csv")
examdata2 <- examData[, c("Exam", "Anxiety", "revise")]
cor(examdata2)

OUTPUT:-

	Exam	Anxiety	revise
Exam	1.0000000	-0.6381787	0.6281441
Anxiety	-0.6381787	1.0000000	-0.8190752
revise	0.6281441	-0.8190752	1.0000000

Interpretation:-

provides a matrix of the correlation coefficients for the three variables. Each variable is perfectly correlated with itself (obviously) and so r = 1 along the diagonal of the table. Exam performance is negatively related to exam anxiety with a Pearson correlation coefficient of r = −.441. This is a reasonably big effect. Exam performance is positively related to the amount of time spent revising, with a coefficient of r = .397, which is also a reasonably big effect. Finally, exam anxiety appears to be negatively related to the time spent revising, r = −.709, which is a substantial effect size. In psychological terms, this all means that as anxiety about an exam increases, the percentage mark obtained in that exam decreases. Conversely, as the amount of time revising increases, the percentage obtained in the exam increases. Finally, as revision time increases, the student's anxiety about the exam decreases. So there is a complex interrelationship between the three variables.

R^2:-

```
> examdata=read.csv("C:\\Users\\aadmin\\Desktop\\examdata.csv")
> examdata2 <- examdata[, c("Exam", "Anxiety", "revise")]
> cor(examdata2)^2     #coefficient of determination
              Exam   Anxiety    revise
Exam     1.0000000 0.4072769 0.3945650
Anxiety  0.4072769 1.0000000 0.6708793
revise   0.3945650 0.6708793 1.0000000
```

Interpretation:-

*Coefficient a step further by squaring it. The correlation coefficient squared (known as the **coefficient of determination**, R^2) is a measure of the amount of variability in one variable that is shared by the other. From the above we may look at the relationship between exam anxiety and exam performance. Exam performances vary from person to person because of any number of factors (different ability, different levels of preparation and so on). then we would have an estimate of how much variability exists in exam performances. We can then use* R^2 *to tell us how much of this variability is shared by exam anxiety. These two variables had a correlation of -0.6381787 and so the value of R^2 will be (-0.6381787) ^2 = 0.4072721.This value tells us how much of the variability in exam performance is shared by exam anxiety.*

If we convert this value into a percentage (multiply by 100) we can say that exam anxiety shares 40.7%1of the variability in exam performance. So, although exam anxiety was highly correlated with exam performance, it can account for only 40.7% of variation in exam scores. To put this value into perspective, this leaves 59.3 % of the variability still to be accounted for by other variables

Practice problem:-

Example: The following table gives the weight (x) (in 1000 lbs.) and highway fuel efficiency (y) (in miles/gallon) for a sample of 13 cars.

Vehicle	*X*	*Y*
Chevrolet Camaro	3.545	30
Dodge Neon	2.6	32
Honda Accord	3.245	30
Lincoln Continental	3.93	24
Oldsmobile Aurora	3.995	26
Pontiac Grand Am	3.115	30
Mitsubishi Eclipse	3.235	33
BMW 3-Series	3.225	27
Honda Civic	2.44	37
Toyota Camry	3.24	32

Hyundai Accent	2.29	37
Mazda Protégé	2.5	34
Cadillac DeVille	4.02	26

Find the Correlation between X and Y

2. Find the Correlation between below data

ENJOY	BUY	READ
4	16	6
15	19	13
1	0	1
11	19	13
13	25	12
19	24	11
6	22	7
10	21	8
15	13	12
3	7	4
11	28	15
20	31	14
7	4	7
11	26	14
10	11	9
6	12	5
7	14	7
18	16	12
8	20	10
2	13	6
7	12	9
12	23	13
13	22	9
15	19	13
4	12	9
3	10	5
9	7	7
7	22	8
10	7	8
2	0	2
15	16	7
1	17	6

3	11	9
6	5	9
13	29	15
15	29	11
16	20	9
14	16	7
1	3	2
8	8	10