Frequently Asked Questions

1. Write the formula used to find product moment coefficient of correlation between two variables.

   Answer:

$$r = \frac{Cov(X,Y)}{\sqrt{V(X)V(Y)}} = \frac{Cov(X,Y)}{SD(X)SD(Y)} = \frac{\Sigma(x-\overline{x})(y-\overline{y})}{\sqrt{\Sigma(x-\overline{x})^2\Sigma(y-\overline{y})^2}}$$

2. Write the simplified formula which we use to calculate product moment correlation coefficient.

   Answer:

$$r = \frac{n\Sigma xy - \Sigma x\Sigma y}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$, this is for raw data and

$$r = \frac{N\Sigma fxy - \Sigma fx\Sigma fy}{\sqrt{[N\Sigma fx^2 - (\Sigma fx)^2][N\Sigma fy^2 - (\Sigma fy)^2]}}$$ This is for tabulated data.

3. How do we interpret the value of r?

   Answer:

   To interpret the value of r, either we can use probable error or coefficient of determination.

4. How do we interpret the data using probable error?

   Answer:

Probable error is given by, $PE = 0.6745\frac{1-r^2}{\sqrt{n}}$

- If the value of r is less than the probable error, there is no evidence for correlation
- If the value of r is more than six times of the probable error, it is significant correlation

5. When do we use probable error and coefficient of determination to interpret the data?

Answer:

Probable error is used to interpret the data only when the number of observations is large. Otherwise it will give misleading results.

Coefficient of determination can be always used to interpret r.

6. How do we interpret the data using coefficient of determination?

   Answer:

After finding coefficient of correlation we find coefficient of determination r^2. Then we convert to percentage and this percentage of variation is explained by the one variable in the other.

7. When we have bigger values of the variable, how do we calculate the coefficient of correlation?
   Answer:
   If the values taken by the variable are very large we can change the origin and scale. But the coefficient of correlation is independent of origin and scale. we can find same formula for calculating the correlation between the original values.

8. In a bivariate data on X and Y, V(X) = 49, V(Y) = 9 and Cov (X, Y) = -17.5. Find the coefficient of correlation between X and Y.

   Answer:
   The coefficient of correlation is, $r = \frac{Cov(X,Y)}{\sqrt{V(X)V(Y)}} = \frac{-17.5}{\sqrt{(49)(9)}} = 0.833$

9. Variables X and Y are perfectly negatively correlated. Also SD(X) =4.6 and SD(Y) =0.47. Find Cov(X,Y)

   Answer:
   The coefficient of correlation is,

$$r = \frac{Cov(X,Y)}{\sqrt{V(X)V(Y)}} \Rightarrow Cov(X,Y) = rSD(X)SD(Y) = -1 \times 4.6 \times 0.47 = -2.162$$

10. Calculate the Karl Pearson's coefficient of correlation between the following two series regarding the age of Husband and wife in years.

Husband's Age	24	27	28	28	29	30	32	33	35	35	40
Wife's Age	18	20	22	25	22	28	28	30	27	30	22

Answer:
Let x denote the age of Husband and y denote the age of wife

X	**y**	**x^2**	**y^2**	**Xy**
24	18	576	324	432
27	20	729	400	540
28	22	784	484	616
28	25	784	625	700
29	22	841	484	638
30	28	900	784	840
32	28	1024	784	896
33	30	1089	900	990
35	27	1225	729	945

35	30	1225	900	1050
40	22	1600	484	880
341	272	10777	6898	8527

The coefficient of correlation between x and y is,

$$r = \frac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

$$r = \frac{11(8527) - (341)(272)}{\sqrt{[11(10777) - (341)^2][11(6898) - (272)^2]}} = 0.504$$

Coefficient of determination $= r^2 = (0.504)^2 = 0.254$

Hence only 25.4 percent of the variation in one variable is explained by the other. Hence there is weak correlation between the two variables.

11. The following table gives the distribution of total population and those who are wholly or partially blind among them.

Age	No. of persons(in '000)	Blind
0-10	100	55
10-20	60	40
20-30	40	40
30-40	36	40
40-50	24	36
50-60	11	22
60-70	6	18
70-80	3	15

Find out if there is any relation between age and blindness.

Answer:
As population of each age group is different, to facilitate comparison it is required to make the number of blinds per equal population and let it be per 1 lakh.

i.e. for the first age group, 0-10, number of persons is 100 thousand and number if blinds are 55. Therefore number of blind per lakh is given by,

$\frac{55}{100{,}000} \times 100000 = 55$ In the second age group, 10-20, number of persons is 60 thousand and number of blinds are 40. Therefore the number blinds per lakh is given by, $\frac{40}{60{,}000} \times 100000 = 67$

Similarly we calculate for all the age groups and let it be denoted by 'y' and since ages are given in terms of class intervals, we take mid values of the class intervals as x. Now let us construct the following table.

x	y	x^2	y^2	xy
5	55	25	3025	275
15	67	225	4489	1005
25	100	625	10000	2500
35	111	1225	12321	3885
45	150	2025	22500	6750
55	200	3025	40000	11000
65	300	4225	90000	19500
75	500	5625	250000	37500
320	**1483**	**17000**	**432335**	**82415**

The coefficient of correlation between x and y is,

Zones	Area(Sq. Km.)	Population	No. of Deaths.
1	200	40,000	480
2	150	75,000	1200
3	120	72,000	1080
4	80	20,000	280

$$r = \frac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}} = \frac{8(82415) - (320)(1483)}{\sqrt{[8(17000) - (320)^2][8(432335) - (1483)^2]}} = 0.898$$

Coefficient of determination $= r^2 = (0.898)^2 = 0.807$

Hence 80.7 percent of the variation in one variable is explained by the other. Hence there is a strong correlation between the two variables. i.e., as age increases blindness also increases.

12. From the following data find out whether there is any relationship between density of population and death rate.

Answer:
In this question we have to find out the relationship between the density of population and death rate, which are not given directly. First of all we shall obtain the density of population and death rate.
To find these we use the following formulae.
Density of population = Population/Area. And

$$\text{Death rate} = \frac{\text{No. of deaths}}{\text{Population}} \times 1000$$

Now let us denote density of population as x and death rate as y.

To calculate correlation coefficient we construct the following table.

X	y	x^2	y^2	xy
200	12	40000	144	2400
500	16	250000	256	8000
600	15	360000	225	9000
250	14	62500	196	3500
1550	57	712500	821	22900

The coefficient of correlation between x and y is,

$$r = \frac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}} = \frac{4(22900) - (1550)(57)}{\sqrt{[4(712500) - (1550)^2][4(821) - (57)^2]}} \text{ =0.8}$$

21

Coefficient of determination $= r^2 = (0.898)^2 = 0.674$.

Hence 67.4 percent of the variation in one variable is explained by the other. Hence there is correlation between the two variables. i.e., there is positive association between densities of population and death rate.

13. A student while calculating the coefficient of correlation between two variates X and Y from 25 pairs of observations obtained the following constants.
n=25, Σxy=508, Σx=125, Σy=100, Σx^2=650, Σy^2 =460.
It was, however detected later on at the time of checking that he had copied down the two pairs of observations, (x,y)= (8,12) and (6,8) as (6,14) and (8,6). Obtain the correct value of the coefficient of correlation.

Answer:
First we correct the given totals by subtracting the wrong entries and adding the correct entries.

Correct value of Σxy=508-(6.14)-(8.6) + (8.12) + (6.8) = 520

Correct value of Σx=125-6-8+8+6=125

Correct value of Σy=100-14-6+12+8=100

Correct value of $\Sigma x^2 = 650 - 6^2 - 8^2 + 8^2 + 6^2 = 650$

Correct value of $\Sigma y^2 = 460 - 14^2 - 6^2 + 12^2 + 8^2 = 436$

$$r = \frac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}} = \frac{25(520) - 125(100)}{\sqrt{[25(650) - (125)^2][25(436) - (100)^2]}} \text{ =0}$$

.67

14. A student calculates the value of r as 0.7 when the number of observations used is 25 and concludes that r is highly significant. Is he correct?

Answer:

For testing the significance we shall calculate the probable error as follows.

$$PE = 0.6745\frac{1-r^2}{\sqrt{n}} = 0.6745\frac{1-0.7^2}{\sqrt{25}}0.0688$$

Six times PE=6(0.0688) =0.4128, which is less than r. Therefore the student is correct.

15. The following is the joint distribution of age of brides and bride-grooms. Calculate the product moment coefficient of correlation.

Age of bridge-groom	**Age of Bride**				
	18-20	**20-22**	**22-24**	**24-26**	**26-28**
20-23	7	6	1	-	-
23-26	3	8	6	4	8
26-29	1	2	3	8	8
29-32	-	1	1	1	2

Answer:
Let X denotes age of bride-groom and Y denotes age of bride.

Since we have given the class intervals, we consider x as mid-point of the class interval corresponding to X and y as the mid-point of the class interval corresponding to Y.

$$r = \frac{N\Sigma fxy - \Sigma fx \Sigma fy}{\sqrt{[N\Sigma fx^2 - (\Sigma fx)^2][N\Sigma fy^2 - (\Sigma fy)^2]}}$$

Age of bridge-groom	**Age of Bride**					**f**	**x**	**u**	**fu**	**fu^2**	**fuv**
	18-20	20-22	22-24	24-26	26-28						
20-23	7 (0)	6 (0)	1 (0)	-	-	14	21.5	0	0	0	0
23-26	3(0)	8 (8)	6 (12)	4(12)	8(32)	29	24.5	1	29	29	64
18	**N=70**			**88**	**162**	**234**	27.5	2	44	88	12
27							30.5	3	15	45	42
4											

72	**150**					
v	0	1	2	3		
fv	0	17	22	39		
fv^2	0	17	44	117	288	**466**
fuv	0	15	30	69	120	**234**

Further we know that correlation coefficient is independent of change of origin and scale, we can write the formula,

$$r = \frac{N\Sigma fuv - \Sigma fu \Sigma fv}{\sqrt{[N\Sigma fu^2 - (\Sigma fu)^2][N\Sigma fv^2 - (\Sigma fv)^2]}}$$

$$= \frac{70(234) - (88 \times 150)}{\sqrt{[70(162) - (88)^2][70(466) - (150)^2]}} = 0.5271$$

Probable error,

$$PE = 0.6745 \frac{1 - r^2}{\sqrt{n}} = 0.6745 \frac{1 - (0.5271)^2}{\sqrt{70}} = 0.0863$$

Six times probable error is (6x0.0863) =0.5179, which is less than r. Therefore the correlation is significant between age of bride-groom and bride.