Frequently Asked Questions

1. What do you mean by probable error?

Answer:

Once we find the correlation coefficient using product moment method, we need to interpret it. Probable error is one of the tools to interpret the correlation coefficient.

2. Explain probable error.

Answer:

After the calculation of coefficient of correlation, the next thing is to find out the extent to which it is dependable. For this purpose, the probable error of the coefficient of correlation is calculated. If the probable error is added to and subtracted from the coefficient of correlation it would give, two such limits within we can reasonably expect the value of coefficient of correlation to vary. It means that if from the same universe another set of samples were selected based on random sampling, the coefficient between the two variables in this new sample would not fall outside the limits so established.

- 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.

3. Write the formula for finding the probable error.

Answer:

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

Where, r is product moment coefficient of correlation and n is number of pairs of observations.

4. When do we use probable error and coefficient of determination for interpreting the coefficient of correlation?

Answer:

Usually we use coefficient of determination for interpreting the coefficient of correlation. We can use probable error for interpretation only when n is large.

5. Why do we find coefficient of determination.

Answer:

It as an important and useful method of interpreting Coefficient of Correlation, which is the square of Coefficient of Correlation or r^2.

Coefficient of Determination = [Coefficient of correlation]^2

The *coefficient of determination, r^2*, is useful because it gives the proportion of the variance (fluctuation) of one variable that is predictable from the other variable. It is a measure that allows us to determine how certain one can be in making predictions from a certain model/graph. The *coefficient of determination* is such that $0 \le r^2 \le 1$, and denotes the strength of the linear association between *x* and *y*.

$$\text{Coefficient of Determination} = \frac{\text{Explained Variance}}{\text{Total Variance}}$$

6. What do you mean by coefficient of non-determination?

Answer:

Total variance = explained variance + unexplained variance.

Sometimes a Coefficient of Correlation is interpreted by finding out the Unexplained Variance. The ratio of unexplained variance to total variance is called the Coefficient of non-determination

K^2 = Coefficient of non-determination

$$= \frac{\text{unexplained Variance}}{\text{Total Variance}} = 1-r^2$$

7. What is coefficient of alienation?

Answer:

K^2 = Coefficient of non-determination

$$= \frac{\text{unexplained Variance}}{\text{Total Variance}} = 1-r^2$$

The square root of K^2 is called the coefficient of alienation and represented by K =square root ($1-r^2$)

8. Where do we use coefficient of determination?

Answer:

Coefficient of determination is used to

- Interpret the coefficient of correlation
- To test how well the regression line represents the data.

9. Explain coefficient of determination.

Answer:

The coefficient of determination (r^2) is a measure of the proportion of variance of a predicted outcome. With a value of 0 to 1, the coefficient of determination is calculated as the square of the correlation coefficient (r) between the sample and predicted data. The coefficient of determination shows how well a regression model fits the data. Its value represents the percentage of variation that can be explained by the regression equation.

A value of 1 means every point on the regression line fits the data, i.e. the dependent variable can be predicted without error from the independent variable. A value of 0.5 means only half of the variation is explained by the regression.

10. How do you interpret the coefficient of determination?

Answer:

In regression, the r^2 coefficient of determination is a statistical measure of how well the regression line approximates the real data points. An r^2 of 1.0 indicates that the regression line perfectly fits the data.

Values of r^2 outside the range 0 to 1 can occur where it is used to measure the agreement between observed and modelled values and where the "modelled" values are not obtained by linear regression and depending on which formulation of r^2 is used.

11. What is adjusted r^2?

Answer:

In many (but not all) instances where r^2 is used, the predictors are calculated by ordinary least–squares regression: that is, by minimizing SS_{err}. In this case r-squared increases as we increase the number of variables in the model (r^2will not decrease). This illustrates a drawback to one possible use of r^2, where one might try to include more variables in the model until "there is no more improvement". This leads to the alternative approach of looking at the adjusted r^2. The explanation of this statistic is almost the same as r^2 but it penalizes the statistic as extra variables are included in the model.

12. What is generalised r^2?

Answer:

- A generalized coefficient of determination should be consistent with the classical coefficient of determination when both can be computed
- Its value should also be maximized by the maximum likelihood estimation of a model
- It should be, at least asymptotically, independent of the sample size
- Its interpretation should be the proportion of the variation explained by the model
- It should be between 0 and 1, with 0 denoting that model does not explain any variation and 1 denoting that it perfectly explains the observed variation
- It should not have any unit

13. Which are things r^2 do not indicate?

Answer:

r^2 does not indicate whether:

- The independent variables are a true cause of the changes in the dependent variable;
- omitted-variable bias exists;
- The correct regression was used;
- The most appropriate set of independent variables has been chosen;
- There is co-linearity present in the data on the explanatory variables;
- The model might be improved by using transformed versions of the existing set of independent variables.

14. From the data given below, calculate product moment correlation coefficient between heights of fathers and their sons.

Height of fathers (inches)	70	71	65	66	67	68	69
Height of sons (inches)	72	69	67	68	66	69	72

Answer:

Since given data is raw data, we know that the formula for calculating the product moment coefficient of correlation is,

$$r_{XY} = \frac{n\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{\sqrt{[n\Sigma x_i^{\ 2} - (\Sigma x_i)^2][n\Sigma y_i^{\ 2} - (\Sigma y_i)^2]}}$$

Hence, we find the following table.

x	y	x^2	y^2	xy
70	72	4900	5184	5040
71	69	5041	4761	4899
65	67	4225	4489	4355
66	68	4356	4624	4488
67	66	4489	4356	4422
68	69	4624	4761	4692
69	72	4761	5184	4968
476	**483**	**32396**	**33359**	**32864**

Now let us substitute the different values in the formula. Here we have 7 pairs or numbers, hence n=7.

$$r_{XY} = \frac{7(32864) - (476 \times 483)}{\sqrt{[7 \times 32396 - (476)^2][7 \times 33359 - (483)^2]}}$$

=0.668

$r^2 = (0.668)^2 = 0.4462$

Therefore, there is moderate positive relation between the height of father and son.

15. Find out Karl Pearson's correlation coefficient between age and playing habit of the following students.

Age (in years)	No. of students	Regular Players
15	250	200
16	200	150
17	150	100
18	120	48
19	100	30
20	80	12

Answer:

Since given data is raw data, we know that the formula for calculating the product moment coefficient of correlation is,

$$r_{XY} = \frac{n\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{\sqrt{[n\Sigma x_i^{\ 2} - (\Sigma x_i)^2][n\Sigma y_i^{\ 2} - (\Sigma y_i)^2]}}$$

Hence, we find the following table. The second column is obtained by dividing the second column by the third column and then multiplying by 100 and rounding to an integer.

Age (in years)	players per 100	x^2	y^2	xy
15	80	225	6400	1200
16	75	256	5625	1200
17	67	289	4489	1139
18	40	324	1600	720
19	30	361	900	570
20	15	400	225	300
105	**307**	**1855**	**19239**	**5129**

Now, let us substitute the different values in the formula. Here we have 6 pairs or numbers, hence n=6.

$$r_{XY} = \frac{6(5129)-(105\times 307)}{\sqrt{[6\times 1855-(105)^2][6\times 19239-(307)^2]}}$$

=-0.979

$r^2 = (-0.979)^2 = 0.9584$

Since r is negative and r^2 is 0.9584, i.e. there is high negative correlation between age of the students and the playing habits of the students. That is they grow older, they play less or their playing habit reduces.