

CORRELATION

1. Introduction

- Correlation is used to measure and describe a relationship between two variables
- Correlation measure three characteristics of relationship
 - The direction of the relationship
 - Positive
 - It means that when value of one variable increase, the corresponding value of related variable also increases.
 - Negative
 - It means that when value of one variable increase, the corresponding value of related variable decreases.
 - Zero correlation (no correlation)
 - It means that when values of one variable increases or decreases independent to the value of other variable
 - The form of the relationship
 - When value of one variable increases, the corresponding value of related variable increases or decreases until certain value, but beyond that value there may have change not in the same trend or may not have any change at all.
 - The degree of the relationship
 - It measure how strong the relationship between the values of two variables

2. Application of correlation

- Prediction

- If two variables are positively or negatively related to each other, then by knowing the value of one of these variables it is possible to predict the corresponding unknown value of the other variable
- Validity
 - Validity is the measure that a test truly is measuring what it claims to measure.
- Reliability
 - Reliability is the measure whether the test instrument produces the stable, consistent measurements it is used again and again in the same group of students or people.
- Theory verification
 - Theory is a statement that makes a specific prediction about the relationship between two variables
 - This predicted relation can be verified by correlation test

3. Measures of correlation

- Pearson correlation (Pearson product-moment correlation)
 - The Pearson correlation measures the degree and the direction of the linear relationship and is denoted by the letter r (correlation coefficient)
 - $r = (\text{degree to which } x \text{ and } y \text{ vary together}) / (\text{degree to which } x \text{ and } y \text{ vary separately})$
 - $r = (\text{covariability of } x \text{ and } y) / (\text{variability of } x \text{ and } y \text{ separately})$
 - $r = (SP) / \sqrt{(SS_x SS_y)}$
 - $SP = \text{sum of product of deviation} = \sum xy - [(\sum x \sum y) / n]$
 - $SS_x = \text{sum of squared deviation of } x = \sum xx - [(\sum x \sum x) / n]$
 - Or $SS_x = \sum x^2 - (\sum x)^2 / n$
 - $SS_y = \text{sum of squared deviation of } y = \sum yy - [(\sum y \sum y) / n]$
 - Or $SS_y = \sum y^2 - (\sum y)^2 / n$
- Spearman correlation (Spearman rank-order correlation)

- It is used when the data are of ordinal variable.
- If it is not then data must be ranked
- Rank order the score separately for each variables with 1 for the smallest score

Case no.	Score for variable 1	Score for variable 2	Ranked for variable 1	Ranked for variable 2
1	3	13	1	2
2	5	14	3	3
3	4	12	2	1
4	6	15	4	4
5	7	16	5	5

- If there are same score for more than one respondents the final rank for the respondents will be the average of the ranks

Respondent no	Score	Rank	Final rank
1	3	2	2.5
2	5	5	5
3	2	1	1
4	3	3	2.5
5	4	4	4

- The equitation for the spearman calculation
- $r_s = 1 - (6\sum D^2)/[N(N^2-1)]$
 - N is the number of pair (xy)
 - D is the difference between each pair (x - y)
- After calculating the value of r or r_s , this is to be compared with the critical value in the correlation table to decide whether there is significant correlation between the variables.

- Calculated value → tabulated value (significant correlation)
- For one-tailed test $df = n-1$ and for two-tailed test $df = n-2$
 - $df =$ degree of freedom
- coefficient of determination
 - this is squared correlation coefficient
 - it measures the percentage of variation shared between the two variables
 - $r = 0.40$
 - $r^2 = 0.16$ i.e. 16%
- Point to be remembered
 - Correlation is not causation
 - Correlation is affected by the range of data
 - Correlation is affected by the outliers

4. Hypothesis tests with the Pearson correlation

- Two-tailed
 - $H_0 = \rho = 0$ (no correlation)
 - $H_A = \rho \neq 0$ (there is correlation)
- One-tailed
 - $H_0 = \rho \leq 0$ (there is no positive correlation)
 - $H_A = \rho > 0$ (there is positive correlation)
- Reporting correlation
 - $r = 0.65$, $n = 30$, $p\text{-value} < 0.01$, one tail or two tail,
 - $r^2 =$ coefficient of determination

5. Summary

- Correlation is a statistical test to assess the relation between two variables
- Relation can be positive or negative
- Two method of test are Pearson and Spearman methods

- Test is used in prediction of relationship testing validity and reliability and verifying theories
- Can be calculated manually using different formulas or using computer statistical package like SPSS
- Correlation does not say about cause and effect relationship
- The correlation coefficient is influenced by the outliers and or range of data under analysis

REGRESSION

1. The statistical technique that is used to determine the best fitting straight line for any set of continuous data set is regression and the line obtained is the regression line.

- Correlation measures if there is any linear relationship between two variables, e.g. liters of oil burn – kilometers of distance runs by the cars
- The line makes it easier to see the relationship
- The line provides a simplified description of the relationship
- The line can be used for prediction, if the value of one variable is known what will be the corresponding value of other variable
- The value is not the actual value if it is measured from the correlation line
- To find out the most appropriate predictable value a regression equation is necessary

2. Equation used in regression

- $Y = bX + a$
 - Where b and a are constant
 - b is called the slope ($b = SP/SS_x$)
 - a is the Y intercept ($a = Y_{\text{mean}} - bX_{\text{mean}}$)
 - this equation determines the change in Y for the change in X

3. The names of the variables on the X and Y axis vary according to the field of application. Some of the common usages are

- X-axis → independent, predictor, carrier, input
- Y-axis → dependent, predicted, response, output