Kenya Medical Training College
Department of Clinical Medicine
Year Two Semester One
Measures of Relationship
19th November 2020

Willis J. Opalla

Learning Objective

- To demonstrate understanding of the knowledge of measures of relationship apply it health statistics.

Learning Outcomes

- By the end of this session, you should be able to
  1. Define the measures of relationship.
  2. Explain the role of measures of relationship in statistics.
  3. Explain the categories of populations based on the number of variables measured.
  4. Explain the type of relationship that may exist between variables.
  5. Apply the coefficients of correlation in making statistical inferences.

Measures of Relationship

- Definition:
  - These are statistical measures that help understand the relationship between two or more variables in the collected data.

Measures of Relationship

- They comprise
  1. Cross-tabulation
  2. Charles Spearman's coefficient of correlation.
  3. Karl Pearson's coefficient of correlation.
  4. Simple regression equations.
  5. Coefficient of multiple correlation.
  6. Coefficient of partial correlation.
  7. Multiple regression equations.

Population Categories

- Populations can be categorized based on the number of different variables measured from the collected data.
  1. Univariate populations
  2. Bivariate populations
  3. Multivariate populations

Univariate Population

- A population in which only one variable is to measured from the collected variable.

Bivariate Population

- A population from which data on two different variables are obtained.
- If for every measurement of variable X, there is a corresponding value of a variable Y.

Multivariate Population

- A population from which data on more than two measured variables are obtained.
- If for every measurement of variable X, there is a corresponding value of a second variable Y and a corresponding value of a third variable Z or a fourth variable W, etc.

Role of Measures of Relationship

- They offer information on how the measured variables relate with one another.
- Example
  - A study may seek to know whether the number of hours students devote for studies is related with their:

a) Family income

b) Age

c) Gender, etc.

Correlation

- Two questions that should be answered on whether a correlation indicates a *causal relationship* in a bivariate or multivariate population:
  1. Is there an association or correlation between the two or more variables? If yes, then to what degree?
  2. Is there any cause and effect relationship between two variables in case of a bivariate population or between one variable and more than two variables in a multivariate population?

Correlation

- The first question is answered by use of *correlation technique*.
- The second question is answered by the *technique of regression*.

In Case of Bivariate Population

- Correlation can be measured through:
    1. Cross-tabulation.
    2. Charles Spearman's coefficient of correlation.
    3. Karl Pearson's coefficient of correlation.
    4. Simple regression equations to measure a *cause-effect relationship*.
- Simple and multiple regression equations measure cause-effect relationships.

In Case of Multivariate Population

- Correlation can be determined through:
  1. Coefficient of multiple correlation.
  2. Coefficient of partial correlation.
  3. Multiple regression equations, to measure a *cause-effect relationship*.

- Simple and multiple regression equations measure cause-effect relationships.

Cross-Tabulation

- Specifically useful when the data are *nominal*.
- Each variable is classified into two or more categories.
- The variables are further cross-classified into sub-categories.
- Interactions between the sub-categories of the variables are examined:

Cross-Tabulation

- The interactions may be:
  1. Symmetrical relationship
  2. Reciprocal relationship
  3. Asymmetrical relationship.
- Symmetrical Relationship:
  - A relationship in which the two variables vary together, but an assumption is made that neither variable is the result of the other.

Cross-Tabulation

- Reciprocal Relationship
  - A relationship in which two variables mutually influence or reinforce each other.
- Asymmetrical Relationship
  - A relationship that exists if one variable, i.e. the independent variable is responsible for another variable (the dependent variable).

Cross-Tabulation

- Begins with a two-way table that indicates whether there is or *there is not* an interrelationship between variables.
- A third factor can be introduced into the association through cross-classifying the three variables.
- By doing this, it can then be found, e.g. that there is a *conditional relationship* in which factor X appears to affect factor Y only when factor Z is kept constant.
- The statistical correlation found through this approach is a weak one: other methods are applied when data is ordinal, interval or ratio.

Cross-Tabulation

- The following measures are applied when data are ordinal or interval or ratio data:
  1. Charles Spearman's coefficient of correlation.
  2. Karl Pearson's coefficient of correlation (or simple correlation).
- Charles Spearman's coefficient of correlation is also known as rank correlation.

Charles Spearman's Coefficient of Correlation.

- Also known as Rank Correlation.
- Applied in measuring the degree of correlation between two variables in case of ordinal data where ranks are assigned to the different values of variables.
- Its main objective is to determine the extent to which the two sets of ranking are similar or dissimilar.

Charles Spearman's Coefficient of Correlation.

- Spearman's Coefficient of Correlation is denoted by r.

$$r = 1 - \left\{ \frac{6\sum d_i^2}{n(n^2 - 1)} \right\}$$

Where

d_i is difference between ranks of the *ith* pair of the two variables.

n is the number of pairs of observations.

Karl Pearson's Coefficient of Correlation.

- The most widely used measure for the degree of relationship between two variables.
- Also referred to as Simple Correlation or the Product Moment Correlation Coefficient.

Karl Pearson’s Coefficient of Correlation.

- It assumes that:
  a) There is a linear relationship between two variables.
  b) The two variables are causally related, i.e. one variable is independent while the other is dependent.
  c) A large number of independent causes are operating in both variables in order to produce a normal distribution.

Karl Pearson's Coefficient of Correlation.

- Karl Pearson's coefficient of correlation is represented by r.

$$r = \frac{\sum(X_i-\overline{X})(Y_i-\overline{Y}).}{n.\sigma_x.\sigma_y}$$

- Or $$r = \frac{\sum(X_i-\overline{X})(Y_i-\overline{Y}).}{\sqrt{\sum(X_i-\overline{X})^2.\ \sum(Y_i-\overline{Y})^2}.}$$

Karl Pearson's Coefficient of Correlation.

where,

X_i is the *ith* value of variable X.

$\bar{X}$ is the mean of X observations.

Y_i is the *ith* value of variable Y.

$\bar{Y}$ is the mean of Y observations.

n is the number of pairs of observations X and Y

σ_x is the standard deviation of X.

σ_y is the standard deviation of Y.

Karl Pearson's Coefficient of Correlation.

- The value of r lies between + or -1.
- Positive values of r indicate a positive correlation between two variables, i.e. changes in both variables take place in the statement direction.
- Negative values of r indicate negative correlation, i.e. changes in the two variables take place in opposite directions.

Karl Pearson's Coefficient of Correlation.

- A '0' value of r indicates tha there is no association between the two variables.
- When r = +1, then this is a perfect positive correlation.
- When r = -1, then there is a perfect negative correlation, i.e. variations in the independent variable X explain 100% of the variations in the dependent variable Y.

Simple Regression Analysis

- Regression is the determination of a standard relationship between two or more variables.
- In simple regression, the independent variable is the cause of the behavior of the dependent variable.
- Regression can only interpret what exists physically, i.e. there must be a physical way by which independent variable X can affect dependent variable Y.

Simple Regression Analysis

- The basic relationship between X and Y is given by:

  $\hat{Y} = a + bX$

  where,

  $\hat{Y}$ denotes the estimated value of Y for a given value of X.

- This is the *regression equation* of Y on X.
- Also known as the regression line of Y on X when plotted in a graph.

Simple Regression Analysis

- The *regression equation* of Y on X :

  $\hat{Y} = a + bX$

- Means that each unit change in X produces a change of b in Y which is positive for direct and negative for inverse relationships.

  $$b = \frac{\sum x_i\, y_i}{\sum x^2_i}, \quad a = \hat{Y} - b\overline{X}$$

- r can then be calculated by:

  $$r = \frac{b\sqrt{\sum x^2_i}}{\sqrt{\sum y^2_i}}$$

Multiple Correlation and Regression

- This is the analysis for relationship when there are two or more independent variables.
- The equation describing the relationship is known as multiple regression equation.

Multiple Correlation and Regression

- Example
  - In the case of two independent and one dependent variable.
  - In this case, the results are interpreted as:
  - Multiple regression equation assumes the form

$$\hat{Y} = a + b_1 X_1 + b_2 X_2$$

where X_1 and X_2 are two independent variables and $\hat{Y}$ is the dependent variable.

Multiple Correlation and Regression

- In multiple regression analysis, the regression coefficients b_1, b_2 become less realistic as the degree of correlation between the independent variables, X_1 , X_2 increases.
- If there is a high degree of correlation between the independent variables, then a *problem of multicollinearity* arises.
- In that case, only one set of independent variables is used for estimation.

Multiple Correlation and Regression

- There is a difference between the collective effect of two independent variables and the individual effect of each of them.
- The collective effect is determined by the coefficient of multiple correlation,

$$R_{y\text{-}x1x2} = \sqrt{\left[\frac{b_1 \sum x_{1i} y_i + \sum b_2 x_{2i} y_i}{\sum x^2_i}\right]}$$

where, $x_{1i} = (x_{2i} - \overline{x}_1)$, $x_{2i} = (x_{2i} - \overline{x}_2)$, $y_i = (Y_i - \overline{Y})$ and b_1 and b_2 are the regression coefficients.

Coefficient of Partial Correlation

- Measures the relationship between a dependent variable and a particular independent variable by eliminating the effects of other related variables, i.e. holding all other variables constant.
- Each partial coefficient of correlation measures the effect of its independent variable on the dependent variable.

Coefficient of Partial Correlation

- The coefficients of partial correlation are referred to as first order coefficients when one variable is kept constant.
- They are called second order coefficients when two variables are kept constant, etc.

Other Measures

1. Index numbers
   - A number used to measure the level of a given phenomenon as compared with the level of some phenomenon in some standard data
   - For studying the changes in the effect of factors which cannot be measured directly.
2. Time series analysis.
   - A series of observations on a particular phenomenon over a certain duration.
   - Done to identify the strategies for achieving the short and long term goals of a project.

Summary

- Spearman's correlation is a non-parametric technique for measuring the relationship between paired observations of two variables when data is ranked.

References

- Kothari, C. R., (2004) *Research Methodology, Methods and Techniques*, 2nd ed., New Age International Publishers, New Delhi.