Topic for the class-Correlation Statistics

Unit _3 : Title-Descriptive statistics

Date & Time : 2.9.24 11.00 AM – 11.50 AM

Dr. Bhramaramba Ravi

Professor

Department of Computer Science and Engineering

GITAM School of Technology (GST)

Visakhapatnam – 530045

Email: **bravi@gitam.edu**

Unit3-syllabus

- **UNIT 3 Descriptive statistics 9 hours, P - 2 hours**
- Measures of Central Tendency – Measures of Variation – Quartiles and Percentiles –

  Moments – Skewness and Kurtosis. Exploratory Data Analytics Descriptive Statistics – Mean,

  Standard Deviation, Skewness and Kurtosis – Box Plots – Pivot Table – Heat Map – Correlation Statistics – ANOVA, Random variable, Variance, covariance, and correlation- Linear transformations of random variables, Regression.
- https://www.coursera.org/learn/data-visualization-r

Understanding correlation statistics

- Correlation statistics are a set of techniques used to measure and analyze the strength and direction of the relationship between two or more variables.
- The most common methods for assessing correlation include Pearson's correlation coefficient, Spearman's rank correlation coefficient, and Kendall's tau coefficient.
- Each of these methods has its own assumptions and applications.

1. Pearson’s correlation coefficient

Pearson’s correlation coefficient (denoted as r) measures the linear relationship between two continuous variables. The value of r ranges from -1 to +1:

- An r value of +1 indicates a perfect positive linear relationship, meaning that as one variable increases, the other variable also increases.
- An r value of -1 indicates a perfect negative linear relationship, meaning that as one variable increases, the other variable decreases.
- An r value of 0 indicates no linear relationship between the variables.

The formula for calculating Pearson’s correlation coefficient is given by:

$$r=n(\sum xy)-(\sum x)(\sum y)[n\sum x2-(\sum x)2][n\sum y2-(\sum y)2]$$

Where:

- n is the number of pairs,
- x and y are the individual sample points.

2. Spearman’s rank correlation coefficient

Spearman’s rank correlation coefficient (denoted as ρ) assesses how well the relationship between two variables can be described using a monotonic function. Unlike Pearson’s method, Spearman’s does not require the assumption of normality and can be used with ordinal data or non-linear relationships.

The formula for Spearman’s rank correlation is:

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

Spearman's rank correlation coefficient

ρ=1−6∑di2n(n2−1)

Where:

- di is the difference between ranks for each pair,
- n is the number of pairs.

3. Kendall’s Tau coefficient

Kendall’s tau (denoted as τ) is another non-parametric measure that evaluates the strength and direction of association between two ranked variables. It is particularly useful when dealing with small sample sizes or when there are tied ranks.

The formula for Kendall’s tau is:

$$\tau=(C-D)12n(n-1)$$

Where:

- C is the number of concordant pairs,
- D is the number of discordant pairs,
- n is the total number of observations.

4. Applications and importance

- Correlation statistics are widely used in various fields such as psychology, economics, biology, and social sciences to identify relationships among variables. Understanding these correlations helps researchers make predictions, inform policy decisions, and develop theories based on empirical data.

- However, it is crucial to remember that correlation does not imply causation; just because two variables are correlated does not mean that one causes changes in another. This distinction must be made clear when interpreting results from correlation analyses.

- In summary, **correlation statistics provide essential tools for understanding relationships between variables**, allowing researchers to quantify how closely related different factors may be while also highlighting important considerations regarding causality.

THANK YOU