Week 4

Central Tendency, Dispersion, and Their Relationships

Data that is quantitative is described in numerous ways. Central tendency is reflected by the mean, median, and mode. The dispersion expresses the degree to which the data is distributed by the measures of central tendency. The dispersion is demonstrated using the range, variance, deviation, standard deviation, and standard error. The standard deviation measures the spread of values in a distribution. Variance measures the distance of spread or variability from the average. Deviation measures the difference between the observed value and the mean of the variable of interest. Standard deviation is the measure of dispersion or "how spread out" a set of data values are. The standard error (SE) is the estimate of the standard deviation signified by a higher number if the spread is wide and a lower number if the spread is narrow. The sample mean deviates from the actual mean of a population and this represents the standard error of the mean.

The Shape of the Data

The data is often referred to as "following a bell curve" or in some cases, the data does not reflect the bell curve. This is known as the shape or the probability distribution of the bell curve. Two characteristics serve as measures to reflect the shape: kurtosis and skewness. *Kurtosis* reflects the height and width of the peak of the bell curve, which focuses on the extent to which the data points cluster around the "middle of the bell." *Skewness* reflects the extent to which there is an asymmetry of the probability distribution. When the shape of the data looks like a perfect symmetrical bell curve, the data is said to be normally distributed.

The Confidence Interval

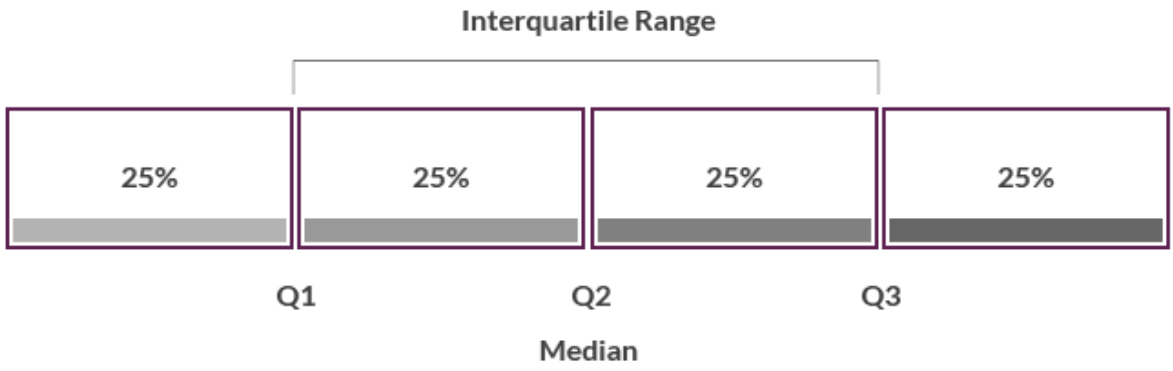
Using the observed data, an estimate derived from statistical analysis is called the confidence interval. The associated confidence level is typically selected by the individual to reflect the probability that the result falling within this range will not be due to chance factors related to the method of sampling. Another way of conceptualizing the confidence interval is the likelihood that a range of values will be found within the actual results. A confidence interval of 95% means that if an analysis was conducted hundreds of times, 95% of the results would fall within the resulting range of values. In actuality, this means that the sample mean can be used as an estimate of the real or "true" population mean.

Interquartile Range

One measure of spread is the interquartile range (IQR). A quartile of a sample or population is simply one of the three values associated with the distribution that is divided by even 4ths. Using Figure 7 as a guide, 25% of the data fall below the first quartile (Q1), 50% of the data fall below the second quartile (Q2, also the median), and 75% of the data fall below the third quartile (Q3).

Figure 6

Interquartile Range

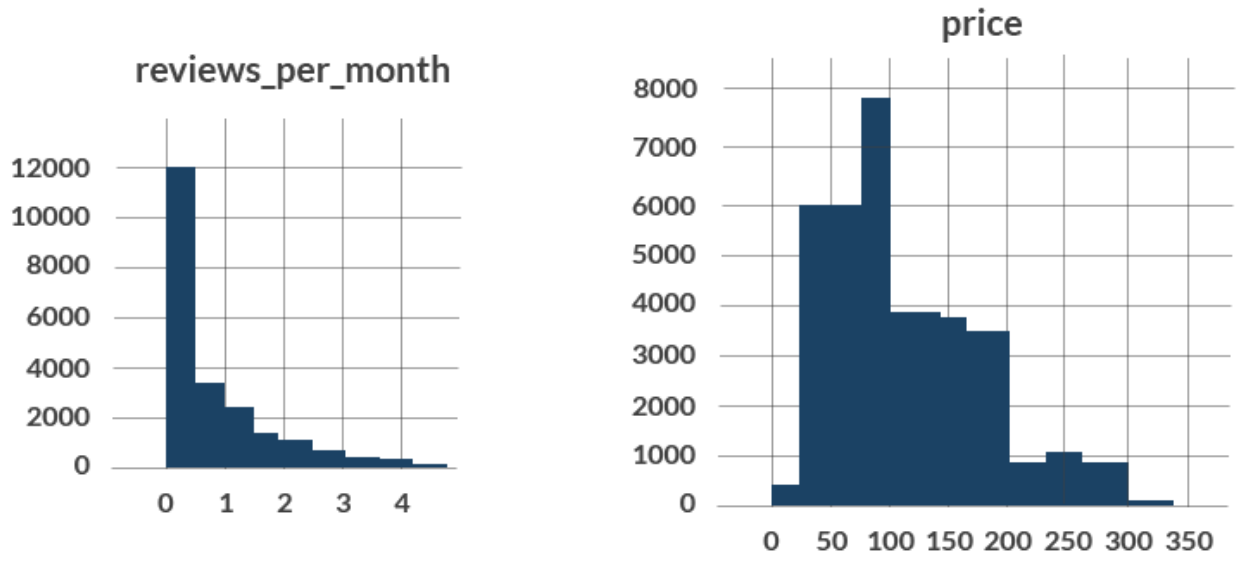


Univariate Analysis

The process of univariate analysis focuses on one feature at a time. The goal of the univariate analysis is to describe the data and determine what patterns are present and this is typically done during the exploratory data analysis phase. When describing the data, the sample distribution can be determined and initial inferences about the population distribution can be made. For categorical data, outliers are often revealed during this process because the tabulation in a frequency table permits the outliers to easily appear. The frequency table can also be used for quantitative variables that do not have very many distinct values. In cases where the quantitative variable has numerous distinct values, the central tendency, spread, skewness, and kurtosis are easily revealed in a histogram.

Figure 7

Histogram

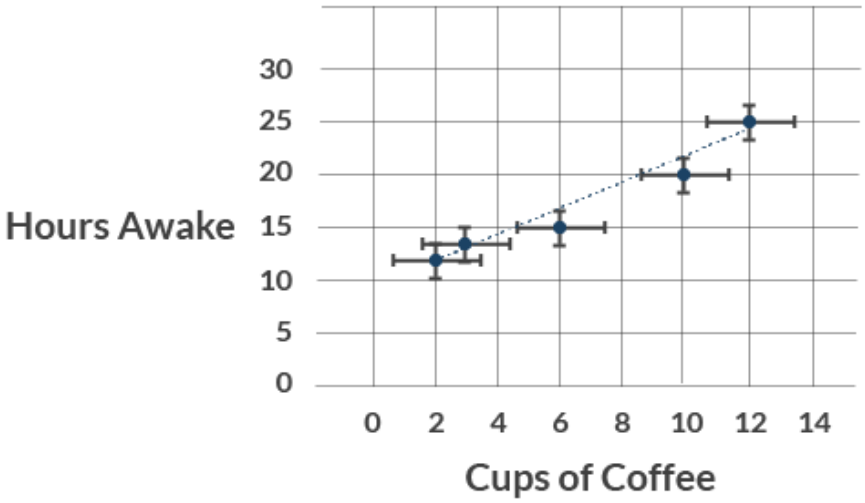


Bivariate Analysis

When the focus is on two different variables and the goal is to determine whether a relationship exists between them, a bivariate analysis is performed. This is typically done during the exploratory data analysis phase. The bivariate analysis examines the relationship between one variable (x) and another variable (y). The goal is to determine the nature and extent of a linear relationship between these variables. For example, the number of cups of coffee could serve as the variable (x) and the number of hours awake could serve as the variable (y). In Figure 8 below, there is a positive slope or upward trend and there appears to be a high correlation between cups of coffee and the number of hours awake. The next step would be to explore even further by investigating the strength of the linear relationship.

Figure 8

Bivariate Analysis



Weekly Resources and Assignments

Review the resources from the Course Resources link, located in the top navigation bar, to prepare for this week's assignments. The resources may include textbook reading assignments, journal articles, websites, links to tools or software, videos, handouts, rubrics, etc.

0 % 0 of 1 topics complete

Week 4 - Assignment: Interpret Summary Statistics and Relationships Between Features

Assignment

Due July 30 at 23:59

This week, the assignment is divided into two parts: code and paper. In the code part of the assignment, you will use perform analyses of the transformed data set associated with your business case scenario using R Studio. You will also document your findings in the paper part of the assignment per the instructions below.

Part 1: Code

- Use the appropriate statistical functions to provide summary statistics for each relevant feature in the data set. Specifically, calculate the measures of central tendency (mean, median, mode), measures of variability (variance, standard deviation, skewness, percentiles (quantiles), and ranges), and the correlation between variables.
- Perform a univariate analysis on each continuous and categorical variable separately using an appropriate visualization (graph).
- Perform a bivariate analysis to find relationships between each of the relevant features in the data set and the target variable of interest using an appropriate visualization (graph) to plot the relationship. Describe and interpret the specific association between variables.

Part 2: Paper

- In a paragraph, describe the general purpose of using summary statistics.
- Describe and interpret the results of the calculation of summary statistics and embed a screenshot of the results of the summary statistical analysis conducted on the dataset.
- Describe the general purpose of the univariate analysis.
- Embed screenshots of each graph for the univariate analyses. Interpret the results by describing the modality, symmetry, distribution range, and mode for each univariate analysis conducted on the dataset. Justify the visualization types used to communicate findings in the univariate analysis.
- Describe the general purpose of the bivariate analysis.
- Embed screenshots of each graph for the bivariate analyses. Interpret the results by describing the association and strength of association between variables, differences between variables, and the significance of those differences for each bivariate analysis conducted on the dataset. Justify the visualization types used to communicate findings in the bivariate analysis.

Length: 3 to 5-page paper, plus complete R code file and dataset (CSV)

References: Include a minimum of 3 scholarly resources.

The completed assignment should address all of the assignment requirements, exhibit evidence of concept knowledge, and demonstrate thoughtful consideration of the content presented in the course. The writing should integrate scholarly resources, reflect academic expectations and current APA standards, and adhere to Northcentral University's Academic Integrity Policy.

When applicable, conduct a Turnitin pre-check and then upload your completed assignment and click the *Submit to Dropbox* button.