1 Data

Definition

Categorical Variables represent data that can be divided into different groups.

Common examples of categorical data include ethnicity, income level, education, age group, and gender. Categories are described by words or letters. Categorical data is much harder to analyze mathematically than numerical data.

Definition

Quantitative Variables represents numerical data from population measurement. Quantitative data can be either **discrete** or **continuous**.

Common examples of quantitative data include height, weight, income, age, and cost. If data set can only take on specific values (e.g. integer values), then the data set is discrete. If the data is not restricted to specific values, then the data set is continuous.

2 Sampling

Gathering the data on an entire population may be virtually impossible due to limitations in time and cost. Instead, a **sample** of the population will be studied. The sample must be representative of the population being sampled in order for any statistical inference to be made. Random sampling is used to to create samples that minimize any bias from the sampling process.

Definition

A **simple random sample (SRS)** is a sampling method in which each member of a population has an equal chance of being selected.

In practice, it may be difficult to achieve a random sample. Care must be taken to eliminate biases that arise in a sampling process. Many data selection methods are not truly random. Random number generators can be powerful tools to add randomization to the sampling process. Two other common types of sampling methods are stratified sampling and cluster sampling. Introductory statistics assumes a simple random sampling process.

3 Histograms

Definition

A **Histogram** is a bar chart representing how many data points are present in each specified interval.

4 Percentiles

5 Box Plots

6 Statistical Mean

Definition

The **mean (arithmetic mean)** is one of the most common measures of center of a data set in statistics. The mean of a data set representing an entire population is called a **population mean** and is denoted μ.

The mean can be calculated by the formula

$$\mu = \frac{1}{N}\sum x_i$$

where the sum is taken over the entire data set and N represents the population size. The mean of a data set representing a sample from a population is called a **sample mean** and is denoted $\bar{x}$. The formuala for sample mean is identical to the formula for population mean

$$\bar{x} = \frac{1}{n}\sum x_i$$

except that the sum is taken over the sample data set and n represents the sample size. For datasets described by histograms, the population mean formula takes the form

$$\mu = \sum (RF_i)x_i$$

where RF_i is the relative frequency of the data point x_i.

7 Skewness