What is Data?

Data are the collection of measurements on observational units.

- **Observational units** are the person/things you want to know about.
- **Variables** are measured on observational units and are any characteristic of a person or thing that can be assigned a number or category.

Data are typically presented in a *tidy* format, where variables are columns and individual observations are rows. Measurements are represented by either discrete data (e.g., blue or yellow) or continuous data (e.g., 12 or 15).

Collection of Data

There are many ways to collect data. The simplest of which is **Simple Random Sampling** (SRS), which implies that each member of the population has an equal probability of being chosen. The **sample size** is simply the number of elements of the sample. A sample is **biased** if the elements of the sample share some property that influences their result

Overview of statistics

Generally, our goal in statistics is to gather information from a sample and then generalize the results from the sample to some larger population of interest. This is done by calculating statistics (numerical properties of the sample) and using them to estimate parameters (numerical properties of the population). This process is called **statistical inference**.

Terminology

The whole purpose of statistics is to understand a set of information. This set of information is from a . . .

> **P**opulation is the complete collection of ALL elements that are of interest for a given problem.
> **P**arameter is the numerical measurement used to describe a population.

The population is often so big that obtaining all information about its elements is either difficult or impossible. So, we work with a more manageable set of data that we obtain from a . . .

> **S**ample is sub-collection of elements drawn from a population. **S**tatistic is the numerical measurement used to describe a sample.

Example

Gallup was interested in why so many millennials are "job hopping." They conducted a poll of 534 millennials in the US workforce. One of the questions asked if they had learned something new in the past 30 days. They were given the following choices: strongly agree, agree, neither, disagree, strongly disagree.

1. We are interested in the proportion of those who strongly agree.
2. What is the population of interest?
3. What is the variable of interest?
4. What is the parameter of interest?

5. What is the observational unit?
6. What is the sample?
7. What is the statistic?
8. What proportion you would expect to select strongly agree by chance?

Article: Why Your Best Millennials Will Leave and How to Keep Them

Descriptive statistics

We now move into looking at ways to describe/summarize data. Two of the things we are most interested in when looking at data are:

Where the data are the data centered .

How spread out the data are.

We will next look at each of these two ideas and finish the chapter by exploring a few ways to visualize data.

Median

Median is defined as middle or average of middle two observations, where the observations are completely ordered.

- When do we use the middle observation to identify the median?
- When do we use the average of middle two observations to identify the median?

Median of 1, 4, 2, 1 is _____

Median of 4, 3, 1, 2, 3 is _____

Mean

The mean/average is defined as the numerical average of all observations. The sample mean is denoted by $\bar{X}$, where

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

The mean of a population is denoted by μ.

Example: the mean of 4, 3, 1, 2, 3 is _____

Example

Consider the following daily minimum temperatures (in Fahrenheit) for the previous 10 days. Calculate the mean and median.

69, 64, 64, 68, 70, 62, 65, 66, 68, 66

Would you prefer the mean or the median for summarizing this data?

Now suppose a cold front occurred on day 11, which had a minimum temperature of 35 degrees Fahrenheit. What is the new mean and median? Which would you prefer for summarizing the data?

Measures of variability

Measures of location are usually considered incomplete without a corresponding measure of variability. The numerical measures of variability we are interested in are:

Variance

Standard deviation

Both numbers provide a measure of how spread out the data are. The larger the number, the more spread out the observations are. Variance is the average squared deviation of data from the mean, and the standard deviation is the square root of the variance.

Population Variance

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}$$

Sample Variance

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} = \frac{\sum_{i=1}^{n} X_i^2 - n\bar{X}}{n-1}$$

Range

$$R = X_{max} - X_{min}$$

Data Visualization

Visualizations are often very useful in helping us understand and draw conclusions from data.

Dot Plot

Histogram

Scatter Plot