

# STAT 515 --- STATISTICAL METHODS

**Statistics:** The science of using data to make decisions and draw conclusions

**Two branches:**

**Descriptive Statistics:**

- The collection and presentation (through graphical and numerical methods) of data
- Tries to look for patterns in data, summarize information

**Inferential Statistics:**

- Drawing conclusions about a large set of individuals based on data gathered on a smaller set
- Example:

**Some Definitions**

**Population:** The complete collection of units (individuals or objects) of interest in a study

**Variable:** A characteristic of an individual that we can measure or observe

**Examples:**

**Sample: A (smaller) subset of individuals chosen from the population**

- **In statistical inference, we use *sample data* (the values of the variables for the sampled individuals) to make some conclusion (e.g., an estimate or prediction) about the population.**

**Example:**

**How reliable is this generalization to the population?  
For inference to be useful, we need some genuine  
measure of its reliability.**

**Types of Data:**

**Quantitative (Numerical) Data: Measurements recorded on a natural numerical scale (can perform mathematical operations on data).**

**Qualitative (Categorical) Data: Measurements classified into one of several categories.**

## **Examples:**

### **Sources of Data:**

- **Published Source:** Many government, business, financial, and sports statistics are collected and archived in publications or online.
- **Designed Experiment:** Researcher imposes a treatment on individuals, then observes responses. (Researcher maintains strict control --- often in lab setting.)

#### **Example:**

- **Surveys:** Researcher selects sample of individuals and records their responses to questions

#### **Example:**

- **Observational Study:** Researcher observes individuals and measures variables, but has no control over process being observed.

#### **Example:**

**Typically, an experiment is better for establishing cause/effect between two variables, but it's not always practical or possible.**

**Regardless of the type of study, we must ensure that we have a representative sample, one that has similar characteristics to the population.**

**The best kind of sample is a *simple random sample* (every subset has an equal chance of being selected)**

**Standard statistical methods assume the data are a random sample from the population.**

## **Methods for Describing Data Sets**

**Important Principle in Statistics: Data Reduction**

**Example:**

- **List of 100 numbers would be confusing, not informative**
- **Need to reduce data to a reasonable summary of information.**

**Two ways:**

- **Graphs and Plots**
- **Numerical Statistics**

### **Describing Qualitative Data**

- **Data are categorized into classes**
- **The number of observations (data values) in a class is the class frequency**
- **Class Relative Frequency = class frequency /  $n$**
- **The CRF's of all classes add up to 1**

**Example:**

### **Graphical Displays:**

**Bar graph:** Height of bars indicates frequencies for each category (see p. 30)

**Pie chart:** Area of “pie slices” indicates relative frequency for each class (see p. 30)

### **Describing Quantitative Data**

To detect and summarize patterns in a set of numerical data, we use:

- **Dot Plots:** these represent each data value with a dot along a numerical scale. When data values repeat, the dots pile up vertically at that value.
- **Stem-and-leaf Display** (good for small data sets): Separate each number in a data set into a stem and a leaf (usually the last digit). There is a column of all the stems in the data set, and at each stem, the corresponding leaf digits line up to the right.

**Example:**

See p. 39 for another example.

- **Histogram**: Numerical data values are grouped into measurement classes, defined by equal-length intervals along the numerical scale.

**Like a bar graph, a histogram is a plot with the measurement classes on the horizontal axis and the class frequencies (or relative frequencies) on the vertical axis.**

**For each measurement class, the height of the bar gives the frequency (or RF) of that class in the data.**

**Example:**

**Guidelines for Selecting Measurement Class Intervals:**

- **Use intervals of equal width**
- **Each data value must belong to exactly one class**
- **Commonly, between 5 and 12 classes are used**

**Note: Different choices of Class Intervals (in position and number) may produce different-looking histograms.**

**Most often, we use software to help choose intervals and create histograms.**

**Histograms don't show individual measurement values (stem-and-leaf displays and dot plots do).**

**But for large data sets, histograms give a cleaner, simpler picture of the data.**

### **Summation Notation**

**In statistics, we customarily denote our data values as  $x_1, x_2, \dots, x_n$ . ( $n$  is the total number of observations.)**

**The sum of a set of numbers is denoted with  $\Sigma$ .**

**That is,  $x_1 + x_2 + \dots + x_n =$**

**Or, we might sum the squared observations:  
 $x_1^2 + x_2^2 + \dots + x_n^2 =$**

**Example:**