What is Statistics?

- A whole **subject** or **discipline**
- The **methods** used to collect, process and/or interpret data
- **Collections of data** gathered by those methods
- **Specially calculated figures** (e.g. averages) to characterize collection

---

What is Statistics?

- A student in a class offered by the PSU **Statistics** department uses **statistics** (statistical methods) to interpret **statistics** (data) about the cost of a 1 bedroom apartment in State College, and he/she may summarize finding by quoting a **statistics** of ‘average price per 10 apartments’ in various locations of State College.

---

What is Statistics?

- The objective of **descriptive statistics methods** is to summarize a set of observations
- The objective of **inferential statistics methods** is to make inferences (predictions, decisions) about **population** based on information contained in a **sample**, and to quantify the level of uncertainty in our decisions.
Observational Units

- Observational units are entities whose characteristics we measure.
- Synonyms: “case” or “subject.”
- In the social sciences, observational units are often people or groups of people. In the life sciences, observational units might be animals, bacterial colonies, etc.

Observational Units

- In the physical sciences, observational units might be "samples" of some chemical substance or some manufactured good.
- It is a bad habit to use the word “sample” for each observational unit; in standard statistical terminology the whole group of observational units are referred to as "the sample."

Population and Sample

- **Population**: the entire collection of units about which we would like information (e.g. all 1bd apartments in State College)
- **Sample**: the collection of units we actually measure (e.g. 10 1bd apartments in SC)
- **Parameter**: the true value we hope to obtain (e.g. true average cost of 1bd apt. in SC)
- **Statistic**: an estimate of the parameter based on observed information in the sample (e.g. average rice of 10 sampled apartments)
- Parameters are generally unknown so we estimate them with sample statistics

Random Variables (RVs)

- Random variables are characteristics of the observational units which can have different possible values (this is the practical, not the statistical definition)
- **Types**
  - **Quantitative** (numerical, measurement) variables represent an amount or quantity of something (e.g. time spent waiting for the bus)
  - **Qualitative** (categorical) variables represent things that can be categorized (e.g. the colors of the cars that pass while you wait for the bus)
- Letters like X or Y represent random variables if its value is not known before the experiment is run.
Quantitative: Discrete vs. Continuous

- **Discrete** random variables can only take on values from a countable set of numbers such as the integers or some subset of integers. (Usually, they can't be fractions.)

- **Continuous** random variables can take on any real number in some interval. (They can be fractions.)

- Note: We consider variables like height to be continuous even though we can only measure them in discrete units (e.g. millimeters).

Categorical: Nominal vs Ordinal

- **Nominal** (unordered) random variables have categories where order doesn’t matter (e.g. gender, ethnic background, religious affiliation, ...)

- **Ordinal** (ordered) random variables have ordered categories (e.g. grade levels, income levels, school levels,...)

Explanatory vs. Response Variable

- Explanatory variable attempts to explain (or is purported to cause) differences in a response variable (or outcome variable).

- E.g. homework scores and exam scores can be explanatory variables for the final grade

Types of Studies

- **Randomized Experiment**: we create differences in the explanatory variable and then examine the results
  - The investigators applies one or more manipulations (i.e. treatments) to the experimental subjects
  - Subjects are randomly assigned to treatments

- **Observational Study**: we observe differences in the explanatory variables
Experiments vs. Observations

- Experiments have:
  - Control groups
  - Placebo
  - Blinding
- Can make **causal inference** from randomized experiments not observational studies
- Confounding

In-class exercise

- CNN poll

---

How do we look at data?

- Exploratory Data Analysis (EDA) is description of data, typically in terms of:
  1. Measure of central location
  2. Measure of spread (dispersion)
  3. Shape of distribution
- Non-graphical and graphical EDA

---

Measures of Central Tendency

- Measures of central tendency are some of the most basic and useful statistical functions. They summarize a sample or population by a single typical value.
- The two most commonly used measures of central tendency for numerical data are the mean and the median.
- **Mean:** The average of all data points
- **Median:** The data point where half of the data lies above and half below it
- **Mode:** The most common value in the data
Mean

- Synonyms: average, arithmetic mean
- Gives an "expected value" (not literally)

- The sample mean, written as $\bar{X}$, equals the sum of observations divided by the size of the sample.
  \[ \bar{X} = \frac{\sum X_i}{N} \]

- The population mean, written as $\mu$, is analogous to the sample mean, but for the whole population.

Median

- Also known as 50th percentile

- The sample median is the middle number (or the arithmetic mean of the two middle numbers in the case of an even number of observations) when the observations are written out in order.

- The population median is the 50th percentile in the whole population.

Example

- You are looking for an apartment in Sate College. At the end of spring semester you talked to 15 people who told you how much rent they pay for a 1 bedroom:

  - $280, $320, $330, $340
  - $370, $375, $380, $380
  - $380, $390, $420, $420, $430

- Mean: $3505/15 = $367
- Median: $375, seven value above and below
- Mode: $380, it appears three times

Mean vs. Median

**Differences**

- The mean is somewhat more "mathematically tractable" (works better with some statistical procedures)

- The median is more resistant to outliers

- The median and mean have slightly different interpretations

**Similarities**

- Both tell about where the "typical" or "central" value in a distribution is found.

- For a symmetric distribution such as the normal distribution, the mean and the median are the same number.
Example of Resistance to Outliers

The mean of 3, 4, 6, 7, 8, 10, 15 is about 7.57.
The mean of 3, 4, 6, 7, 8, 10, 150 is about 26.86.
The median of either data set is 7.

Most statisticians would say that in this situation the median is the better measure of central tendency to use. (Of course, it is best to report both.)