

## Basic Statistics Definitions:

**Statistics** – Practice or science of collecting and analyzing numerical data

**Data** – Values collected by direct or indirect observation

**Population** – Complete set of all observations in existence

**Sample** – Slice of population meant to represent, as accurately as possible, that population

**Measure** – Measurement of population/sample, an example would be some “score” (a.k.a. an observation)

**Hypothesis** – Educated guess about what’s going on

**Skew** – Not symmetrical, crooked or uneven

**Impute** – To fill in missing values

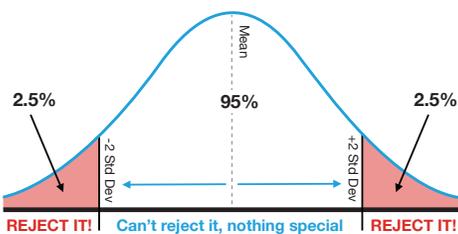
**Type I Error (false positive)** – In hypothesis testing, when you incorrectly reject Null Hypothesis

**Type II Error (false negative)** – In hypothesis testing, when you incorrectly fail to reject Null Hypothesis

## Is My Data Special?

### Null Hypothesis in Layman’s Terms:

*There is nothing different, or special, about this data*



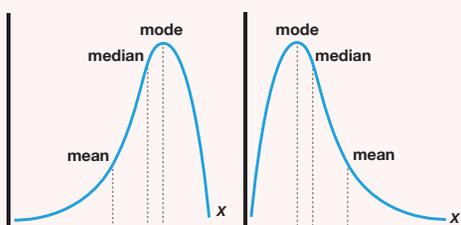
- **Best used when you need to know if your data is different or somehow special**
- Always start out assuming Null Hypothesis is **TRUE**
- Goal is to either “reject” or “fail to reject” Null Hypothesis
- If **FAIL TO REJECT** Null Hypothesis then there is nothing really different about the data
- If **REJECT** Null Hypothesis then we are confident that what we see is different or special
- On curve above, can only say that an observation is different/special if it falls in either of shaded regions (called “tails”)
- The tails are 2 Standard Deviations away from (either above or below) the Mean
- Assumes dealing with a normal distribution!

See Hazards!

**Big takeaway:** If your data falls within +/- 2 Standard Deviations of Mean then its probably not all that different. If your data falls outside those boundaries then it is most likely something to take note of.

### Caution Hazard

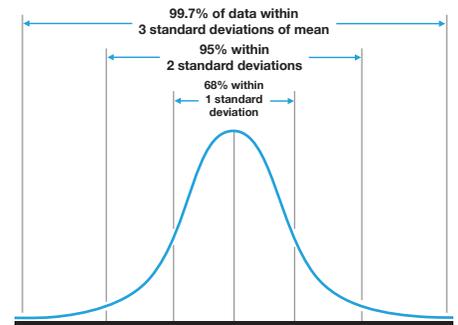
## Skewed Distributions...



Not all data is normally distributed... and when your data is not normally distributed, all those helpful characteristics of a normal distribution no longer apply! For instance Hypothesis testing limits will change, Mean & Median will shift, and most statistical models (think regression) rely heavily on assumption that your data is normally distributed!

## A Normal Distribution:

- A.K.A. “Bell Curve”
- Way to visualize how volume of a population is distributed based on some measurement
- Largest volume is packed around middle
- Volume curves down towards zero to left and right
- Symmetrical around middle
- Interesting Fact: The Mean, Median, and Mode are all the same and at the exact center



## How We Describe Things...

*(Measures of Central Tendency)*

**Mean** – Also called “Average”, probably the most popular statistic, calculated as sum of all values divided by number of values

**Median** – Value at center

**Mode** – Value that occurs most

**Standard Deviation** – Measurement relative to mean, so a measure of how far a value is away from the mean. The further a value is from the mean the more unique... and perhaps interesting... it becomes.

Make sure to review Hazards! section regarding skewed distributions

**Big takeaway:** Most measurements of a normally distributed population will be centered around the middle.

**Why you care:** If population is “normally distributed” then we can use a bunch of useful characteristics to help describe it.

## Sampling

### Good Sampling Rule of Thumb:

- Consider sampling when population you’re working with is **too big to handle**
- Aim is to get a good representative for actual population
- Generally the bigger the sample the better, but a simple tip is:
  - At minimum your sample size should be 100
  - At maximum your sample size should be 10% or 1000, whichever is smaller
- Keep bias out of it by ensuring a **RANDOM** sample!

### Some Sampling Methods:

- **Simple Random**  
*(probably the only one you will ever see or use)*
- **Systematic Random**
- **Stratified**
- **Cluster**
- **Multistage**

## Random Numbers

Are an excellent way to create a Simple Random Sample. Most analytical tools (including Excel & Google Sheets) have a random number generator you can use. Just apply a random number to each row, sort in ascending order by the random number then select the top however-many rows.

### Caution Hazard

#### Beware of... **BIAS**

**Bias** can effect both how samples are selected, and also what conclusions you draw from them (i.e. interpretation).

**Selection Bias** – when an individual or observation is more likely to be picked for sampling (in other words, NOT random)

**Observer Bias** – when you subconsciously let your preconceptions influence how you perform your analysis

**Detection Bias** – when something is more likely to be detected in a specific set of observations (e.g. measuring website traffic on Black Friday)

**Funding Bias** – when selection or interpretation favors a financial sponsor

**Extrapolation Bias** – when you assume results of a study describe a larger population than what you originally started with (e.g. assuming a study of college students is a good proxy for entire country)

**Reporting Bias** – when availability of data favors a certain subgroup within true population

**Confirmation Bias** – tend to listen only to information that confirms hypothesis, assumption, or opinion

## Imputing Missing Values...

Missing values are a part of real-life data analysis. But, resist temptation to just fill them in with Mean or Median.

Sometimes this is an OK option, but remember that missing values can be trying to send you a message about some process that you are unaware of (i.e. telling a story).

Also, there are a number of imputation methods out there, be sure to review them thoroughly to see if there are any that better fit your needs/data.

## Confusing Confidence Intervals...

...with probability. 95% confidence just means that 95% of the time the true (population) value will be within the limits.

## Multiple Inference...Faking it ‘till you’re making it

Running a hypothesis test over and over, the same way on the same data, until you get a “significant” result greatly increases chances you will get a false positive (Type I Error) result because... there is always the chance of getting a randomly significant result.

## Thinking that Correlation proves Causation (it doesn’t)

Check out [Probability & Correlation Cheat Sheet](#) for more on this one!

