- **Anecdotal Evidence : Evidence based froma a limited sample size which might not be representative of the population**

First question you should ask : What is the poulation and what is the sample?
Research Question $\longrightarrow$ Population $\longrightarrow$ Sample $\longrightarrow$ Generalze to

Type of variable:

- Numerical: TAkes numerical values (Add, subtract on these)
  - Continuous : E.g height
  - Discrete: E.g. age
- Categorical : Limited number of distinct categories, can be identified with numbers but no aritmeitc operation
  - Ordinal: Have ordered levels , Satisfactory, poor, very poor
  - Regular : Morning person or Afternoon person

When variables show some connection with one anotehr they are called associated or dependent can be neg or postive . Not associated = independent

- Observatoional Study: Collect data without affecting how data arises, "observe"
  - Retrospective: Data collected from past
  - Prospective: Collected throught study
- Experimental Study: Randomly assign subjects to treatment, thus can establish causal connections between the explanatory and observed variables

In an observationsal study it is difficult to conclude, unless you really control the effect of other variables. In experimental study dues to rando assignment this is taken care of.

Confoundry variables: Variables that affect both explanatory and response variables

Why is 'census' not a good idea?:

- Part of the 'census' might not be representative of the population, e.g. Immigrants
- Population is dynamic! You dont taste whole of soup to find out its taste! Expolratory analysis!
- For inference to be valid, the sample should represent populaton. Stir the soup before tasting!

Bias:

- Convenient Sample: Pick up people from your class for study because they are easily accesible
- Non response: Non-random section of pople respond to your survey. Emailing a survey to people who do not have internet connection => No point!

- Voluntary Response: Only people who volunteer to respond, respond. not everyone

The Literary Digest shut down because its sample was not representative of the population, however large the sample was! Sampling :

- Simple Random Smapling: Randoml;y select samples so that each data point has equal probaility to get sampled
- Stratified: Divide population into homogenous groups(strata) and sample from these. Eg. if you want males and females to be equally repreented divide them into males and females and then sample
- Cluster: Divide into non-homogenous cluster, sample the clusters and then sample the data points. For eg.e divide the geograophy into clusters , travel to few clusters only

Principles of experiment design:

- Control: Compare treatment of interest to group
- Randomize: random assignment of subject to treatment
- REplicate; replicate entire study or more samples
- block: Block vairbales that might affect response variables. Divide pro and amateur atheletes into two groups, then assign them independently to treatment and control and then observe.

Blcoking vs explanatory

- Explantory varrables/factors are conditions we can impose on the experimental units
- Blocking variables are characteristics that experimental units come with that we would like to control
- placebo: fake treamtent
- placebo effect:eoperiments show effect just because they think they are undergoing treatment
- blinding: subjects dont know whether they are in control or sample
- double blind: nor the experimentalist nor subejct knows the groups

RAndom sampling" Randomy select subjects from sample Random assignment: randomly assign subjects to treatement and cotnrol, thus removing the chance of difference

First you sample randomly from the population and then randomy assing them to control and treamtnet groups

RS,RA => Causal generalisable [IDEAL experiment] NO RS, RA => Causal not generalizable [Most Experiments] RS, NO RA => Not causal but generalizable [MOST Observations] no Rs, no ra => neither causal nor gneralisable, ONLY Correlational [BAD Observationa]

Scatter plot => Explanatory varibale on X axis, Response on Y. Association might be identified, causalaity cannot.

Identifying relationship

- Direction: positve, negative
- Shape: Linear, curve
- Strenght: Strong, weak
- Outliers

Skewness is determined by the longer tail! Modality: Peak in the distribution, unimodal, bimodal, uniform, multimidal

Boxplot: Median and IQR . Using boxplot, you can sketch back the historgram

Mean < Median : Left skewed distribution Mean > Median: Right skewed distribution Mena Median : Symmetric

Meaures of spread:

- Range= max-min
- Variance = avg squred deviation from mean

$$s^2 = \frac{\sigma(x + i - x)^2}{n - 1} \tag{1}$$

- Std deviation