Data Driven Decisions

* Need statistics
* Statistics – Decision making in the presence of uncertainty
* Important Vocabulary
  + Population – the group of people or things that we want to study
  + Sample – the group from the population that we have to study
  + Individuals – members from the population , , ,…
  + Parameter – the numerical value that I am looking for from the population (usually a proportion or mean ). An unknown value that describes our population.
  + Statistic – the numerical value that I have from the sample (usually a proportion or a mean ). A known value that describes our sample. It may or may not describe our population as well.
  + Parameter is to population as statist is to sample.
  + Census – when the sample is equal to the population.
  + Variable – a characteristic for the individuals in the population.

Example: Suppose we want to study the average life expectancy of mice. Our **population** is all mice. We cannot study all mice, we therefore study some mice. These mice that we are actually studying is our **sample**. We then observe these mice from birth to death and take the average life span of the mice in our sample. The life span is a **variable**. This average is our **statistic**. The **parameter** would be the average life span of all mice. Since we cannot find the **parameter**, we use the **statistic** to estimate the **parameter** of the average life expectancy of all mice.

Example: Suppose we want to study if most dogs growl at strangers. Our **population** is all dogs. We cannot study all dogs, we therefore study some dogs. These dogs that we are actually studying is our **sample**. We then observe the dogs in our sample to see if they growl. Whether the dog growled or not is a variable. We then find the proportion that growl. This proportion is our statistic. The **parameter** would be the proportion of all dogs who growl at strangers. Since we cannot find the **parameter**, we use the **statistic** to estimate the **parameter** of the proportion of all dogs who growl at strangers.

Example: Suppose we want to study how students in Professor X’s class perform in STAT101. From the Registrar, we get all the grades for any student who has taken STAT101 from Professor X. The grade for the student is the **variable**. We take the average grade. This would be an example of a **census** since we have all the measurements from each individual in the population.

* Sampling
  + Convenience Sampling – Sampling technic where the sample is chosen in a way that is easy for the collector. Often yields bias results.
    - Example: Standing outside of a sports venue to ask survey questions.
  + Voluntary Response Sampling – sampling technic where the surveyor relies on participants to come forward with their contribution.
    - Example: when you are asked for your feedback on the back of receipts.
  + Random Sampling
    - Any sampling method that incorporates a random component.
    - Examples
      * Stratified Random Sampling
        + Separates the original population into groups or strata, then takes random samples from each strata.
        + Example: A university wants to sample their student body on parking, but wants to be sure that graduate students represent 1/3 of the sample. They would randomly sample from graduate students and then randomly sample undergraduates a sample that is twice as big.
      * Cluster Sampling
        + Population is separated into clusters. Then, random clusters are chosen. Only individuals in these clusters are studied.
        + Usually used to reduce cost
        + Example, want to study dumpster usage in st. louis city. It can be expensive to have someone go through and look at each blocks’ dumpster usage. So, we select 7 random neighborhoods (clusters) to study.
      * Cluster versus stratified
        + In stratified sampling each strata is represented in the sample whereas in cluster sampling only the clusters chosen are represented in the sample.
      * Simple Random Sampling
        + A specific type of random sampling which satisfies the following 2 criteria for a sample of size *n*

Each individual from the population has an equal chance of being chosen for the sample.

Each group of size *n* has an equal chance of being chosen for the sample.

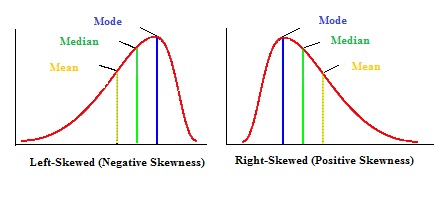
Example: choosing names out of a hat.

Example: assign each individual a random number and then take the first *n* random numbers when sorted in increasing order.

* Questions to ask when looking at data.
  + Who financed acquiring the data?
    - If it is a study on the addictive nature of tobacco and Nicorette financed it, maybe you want to be careful.
  + Who collected the data? Was it independent from those financing it and/or asking the question?
    - If it is a study on the effectiveness of acetaminophen on migraines and Johnson and Johnson ran the study, maybe this raises a red flag.
  + Do they explain how they got the data? What they actually measured and how?
    - Sampling method, measurements, etc.
    - Suppose a study finds that acetaminophen greatly helps with migraine pain with 71% of study participants reporting a significant reduction in pain when using acetaminophen.
      * They studied a sample of 35 women between the ages of 10 and 20, from Springfield, IL who suffer from occasional migraine like pain.
      * They studied 578 patients who have been seen for migraine pain at least once by their primary care physician. They asked individuals to go a month with no pain medication and rate their pain during a migraine. Then, they asked the same individuals to go another month using acetaminophen when they experienced migraine pain and rate their pain 20 minutes after taking the medicine. 71% of patients had a lower average during the month when using acetaminophen.
* Types of Studies
  + Observational Study
    - We do not impose a treatment, but rather observe and gather existing data.
  + Experiment
    - A treatment or treatments is/are imposed and another variable is measured to study the effects of the treatment(s).
    - Placebo – a type of treatment that looks, feels, and tastes like the other treatments. This is to account for the power of suggestion.
* Types of variables
  + Quantitative Variable
    - Variable that is expressed as a number
    - The average would have meaning
    - Example: Age, height, stock price, house price, number of months, etc.
  + Categorical Variable
    - Non numerical variable.
    - Use counts and percentages
    - Examples: genre, color, country, occupation, gender, etc.
* Visualizations
  + for quantitative variables in data
    - Histogram
    - Stem and leaf chart
    - Box plot
  + Visualization for categorical variables in data
    - Bar Chart
    - Pie Chart
  + Variables change over time
    - Line Graph
  + Relationship between 2 quantitative variables
    - Scatterplot
* Summary Statistic
  + For quantitative data
    - Shape, Center, Spread
      * Shape
        + Skewed to the left – bulk of the data is to the right with a tail to the left

Example: Life span

* + - * + Skewed to the right – bulk of the data is to the left with a tail to the right

Example: Income

* + - * + Symmetric – same on left and right

Uniform – all observations are equally distributed

Bimodal – two peaks, one on either side

Normal – single peak in the center

* + - * Symmetric data
        + Mean (for center) and standard deviation (for spread)
      * Non-symmetric data
        + Median (for center) and quartiles/min and max (for spread)
  + For qualitative data
    - Frequency Table

Inference

* Hypothesis testing
* Confidence Interval