Day 4 and 5

An introduction to Statistics

What is Statistics?

- Statistics is all about data.
- Statistics is mathematics used to summarize, analyze, and interpret a group of numbers or observations, but it is only a tool

Classifying Statistics?

- Statistics is classified into two types
    - Descriptive Statistics - describing a lot of numbers using a few numbers
        - Summarise, organize and make sense of data.
            - Measures of central tendency
            - Dispersion
    - Inferential Statistics - describes a facet of stats that deal with making inferences from data.
        - Allow researchers to generalize and infer

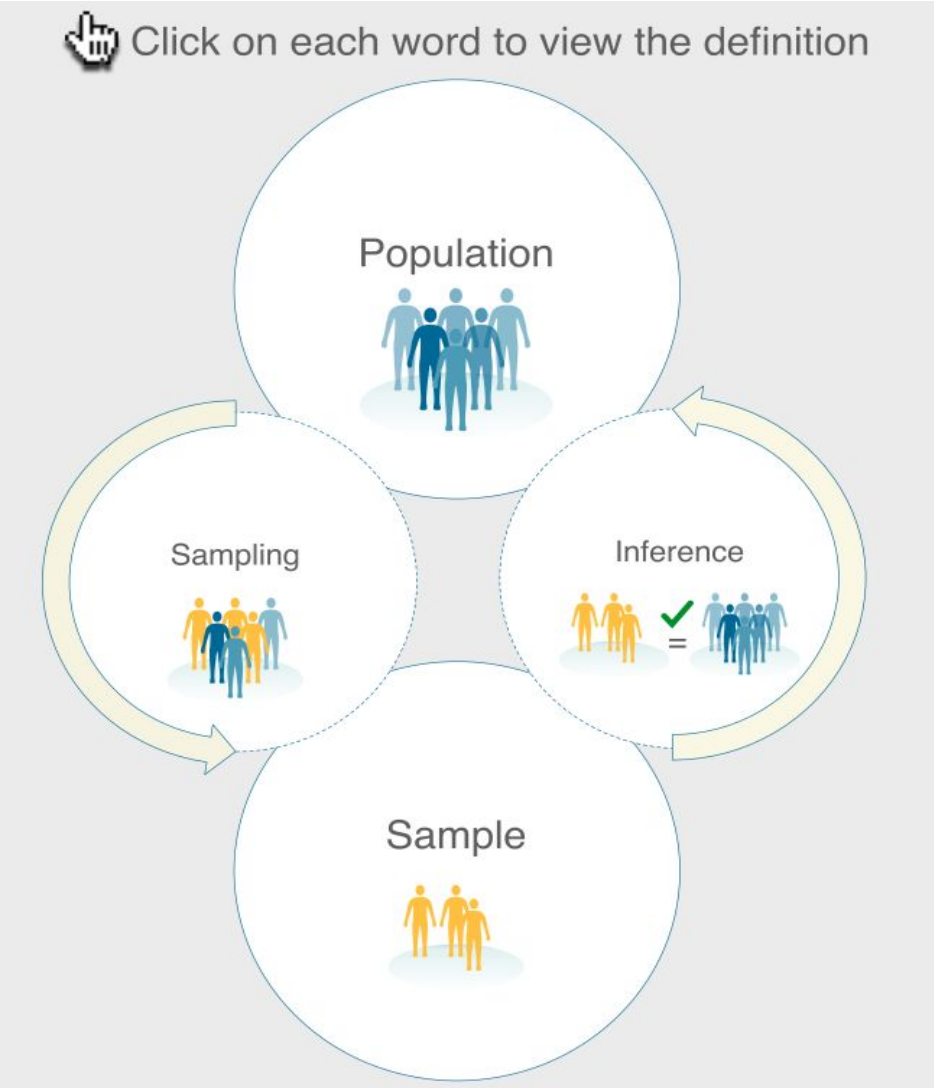
Central Tendency in Statistics?

- Central Tendency
    - The tendency of values in data to revolve around some criteria (mean,mode,median)
    - Used when dealing with random variables.
- Measures of Central Tendency
    - Let's say, we are given 4 3 1 6 1 7
    - Arithmetic Mean
        - Sum of all numbers divided by the number of numbers
        - 4+3+1+6+1+7 = **22/6**
        - **(∑x)/n**
        - Chosen when the data is normally distributed
        - Measures on internal or ration scale
    - Median
        - Looks for the middle number of a sorted sequence of numbers
        - It is a useful measure if the dataset is riddled with outliers, and skews the A.M.
        - ~~1~~ ~~1~~ 3 4 ~~6~~ ~~7~~ = 3+4/2 = **7/2**
        - **(n+1)/2 position is the median.**
        - Chosen when data isn't normally distributed
    - Mode
        - The mode refers to the value that has the highest frequency of occurrence

        - 4+3+1+6+1+7 Most frequent is 1
        - This measure is useful when outliers are present, and they ruin the A.M.

Technical Terms used in Statistics

- Population - A group of interest that researchers want to know more about.
    - Population Parameters - Characteristics of a population.
    - Population Size - (N)
- Sample - A group of individuals or data that are drawn from a population.
    - Sample statistics - characteristics of a sample
    - Sample Size - (n)



Types of measurement

- Continuous
    - Sample space contains a continuous span of real numbers

    - Eg. 6.333333333 is continuous
- Nominal
    - Categorical
        - Numbers / variables that are used to classify into specific groups.
- Ordinal
    - Refers to measurement that classifies on a relative scale.
    - E.g. On 1-10, how bad does this hurt.
- Interval
    - Refers to a scale of measurement that classifies, but also uses 0 as a point of measurement
    - E.g. Fahrenheit
- Ratio
- Qualitative vs Quantitative : Qualitative relates to a verbal description whereas Quantitative relates to a numerical description.

Dispersion (Variability):

- A measure of the spread of scores in a distribution of data.
- Different distributions exist
    - Flat distribution
    - Narrow distribution
- **Range** :
    - Difference between largest and smallest value.
    - Informative for data without outliers
- **Variance :**
    - Measures the average squared distance of the data points from the mean.
    - Represented by σ^2
    - $\sum (x - x')^2/n-1$
    - Degrees of Freedom : n-1 (write in detail)
- **Standard Deviation:**
    - The square root of variance
    - When individual scores are close to mean, the standard deviation (SD) is smaller.
    - The converse holds true
    - Represented by σ
    - $(\sum (x - x')^2/n-1)^{0.5}$

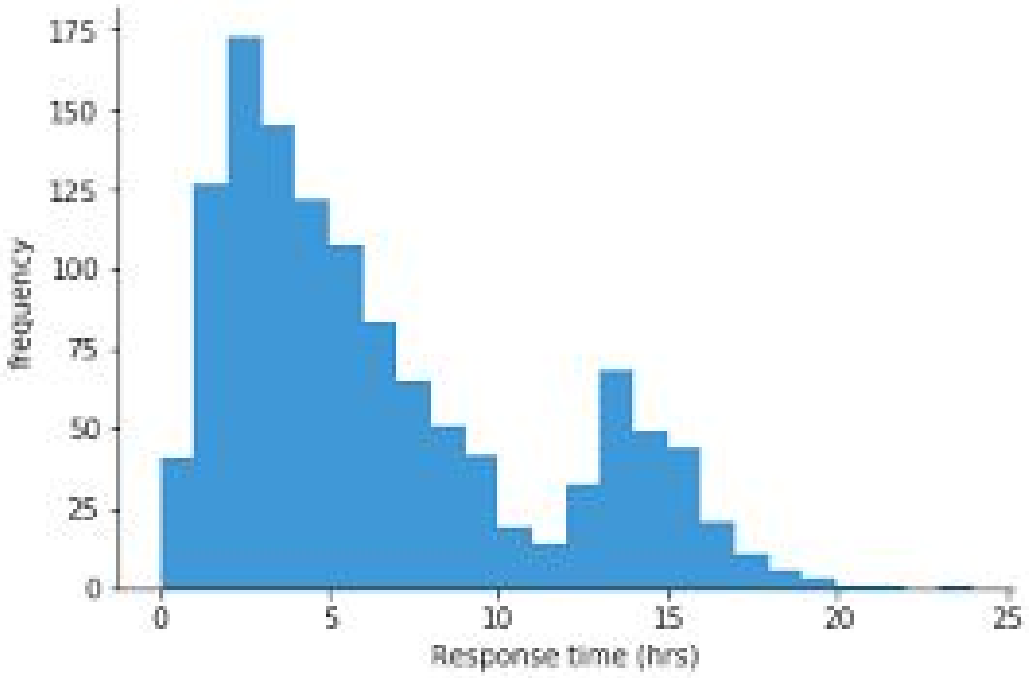
Normal Distribution

- **Probability**
    - The number of times an outcome is likely to occur
    - Varies between 0 and 1
- **Normal Curve:**
    - Occurs when data is symmetrically distributed around mean, median and mode
        - Form of symmetry depends on mean and standard deviation
    - Also called bell-shaped curve

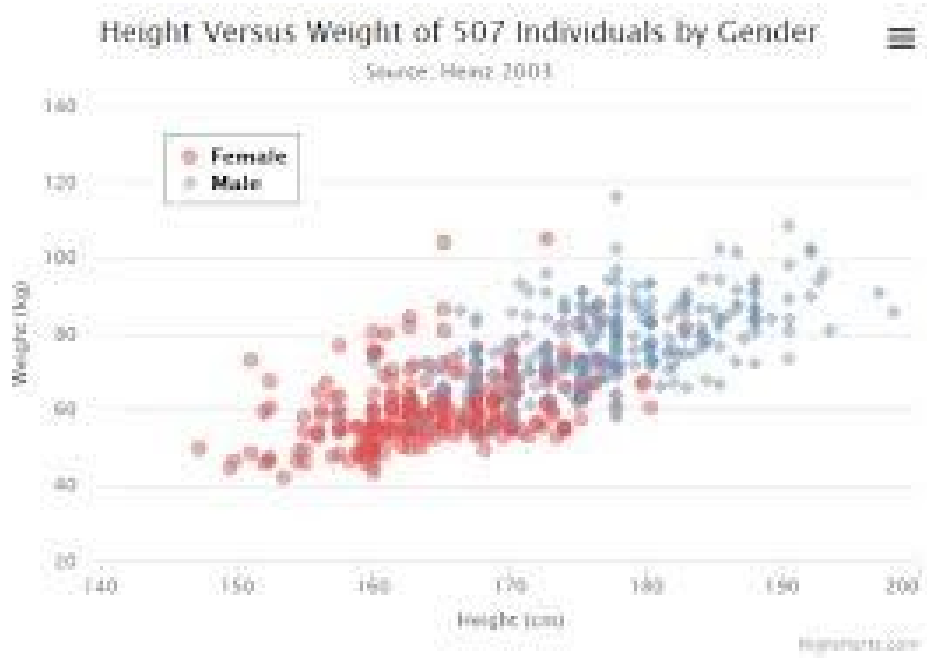
    - Area under curve is 1
    - Tails approach x axis but never touch it
    - **Empirical Rules**
        - 68% lie within one SD of the mean
        - 95% lie within two SDs of the mean
        - 99.7% lie within three SDs of the mean
- **Standard Normal Distribution**

Various Visualizations (Graphs):

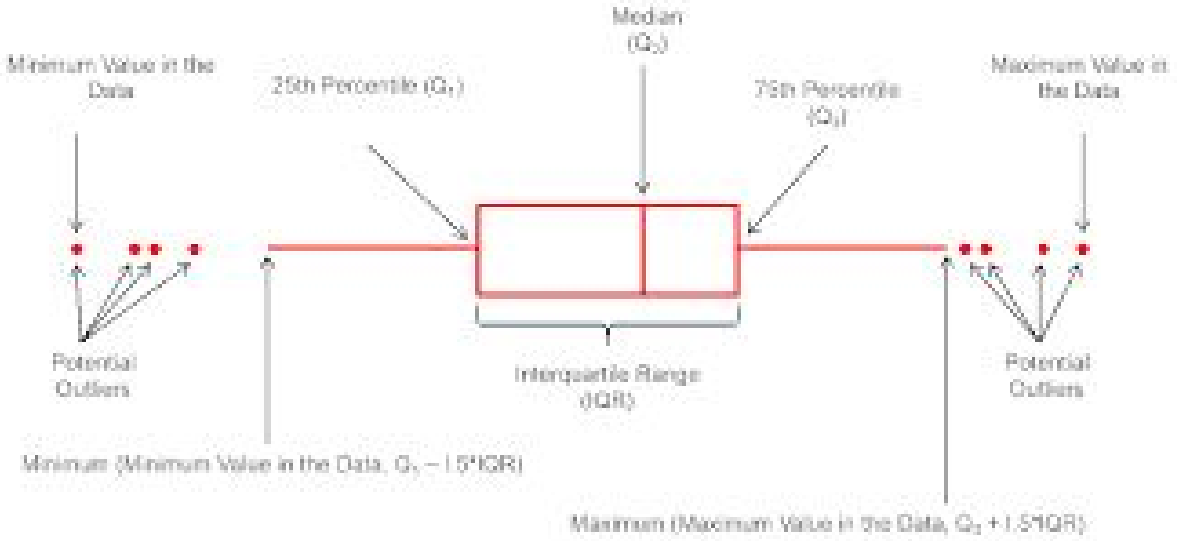
- Used to summarize data graphically
- **Histogram**
    - Used to summarize discrete data (non-continuous)
    - X refers to responses
    - Y refers to frequency



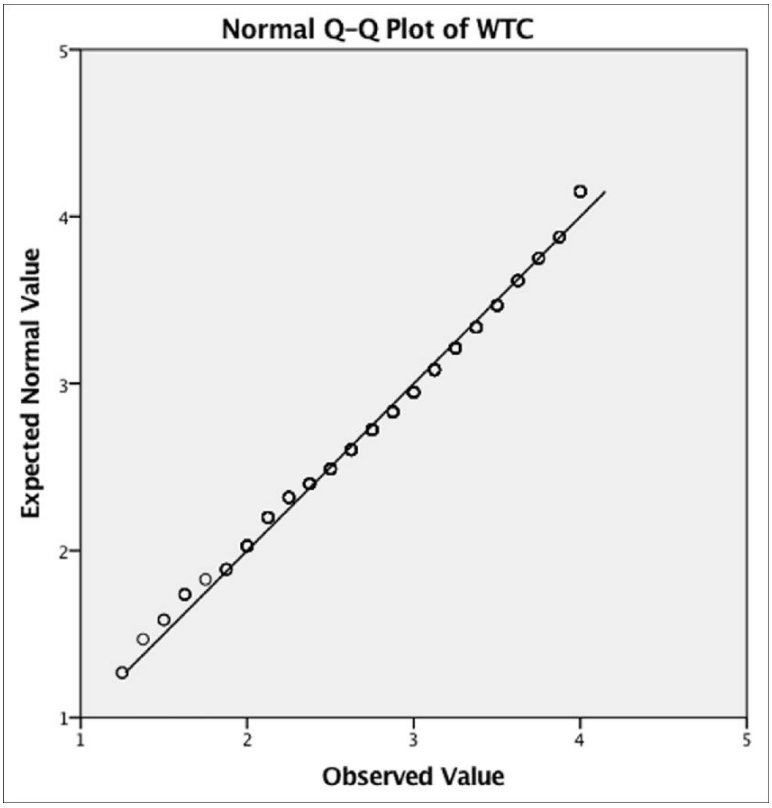
- **Scatter Plot**
    - Used to check linear relationships between two variables
    - Eg. Height vs Weight by Gender



- **Box and Whisker Plot**
  - Good for outlier detection.
  - Tell us where Q1,2,3 are.
  - IQR is calculatble



- **Normal Q-Q Plot**
  - Used to check normality assumption
  - If the data is normal in nature, then the points on the Q-Q plot will fall on a straight line.
  - Eg. Height vs Weight


Normal Q-Q Plot of WTC
Expected Normal Value
Observed Value
1
2
3
4
5