Statistics

- ***Descriptive*:**
  - **Quantitative: 1) continues 2) discrete**

  - **Continues:**

1. **continuous data is the data that can be of any value. Over time, some continuous data can change.**
2. **It may take any numeric value, within a potential value range of finite or infinite.**
3. **The continuous data can be broken down into fractions and decimals, i.e. according to measurement accuracy, it can be significantly subdivided into smaller sections.**
4. **Continuous Data Examples: Measurement of height and weight of a student, Daily temperature measurement of a place, Wind speed measured daily, etc.**
5. **measurable**

- **Discrete:**

1. **Data that can only take on certain values are discrete data.**
2. **These values do not have to be complete numbers, but they are values that are fixed.**
3. **It only contains finite values, the subdivision of which is not possible.**
4. **It includes only those values which are separate and can only be counted in whole numbers or integers, which means that the data cannot be split into fractions or decimals.**

5. **Discrete Data Examples: The number of students in a class, the number of chocolates in a bag, the number of strings on the guitar, the number of fishes in the aquarium, etc.**
6. **Countable**
7. **To consider if we have continuous or discrete data, we should see if we can split our data into smaller and smaller units.**

- **To measure them we use 4 methods :**

1. **Center**
2. **Spread**
3. **Shape**
4. **Outliers**

- **We can divide center into 3 ways:**

1. **Mean:**

- **Is the result of dividing sum of set of numbers on their count .**
- **We can represent it by sigma shape or using x or y bar**

2. **Median:**

- **The middle number of group of sorted numbers**

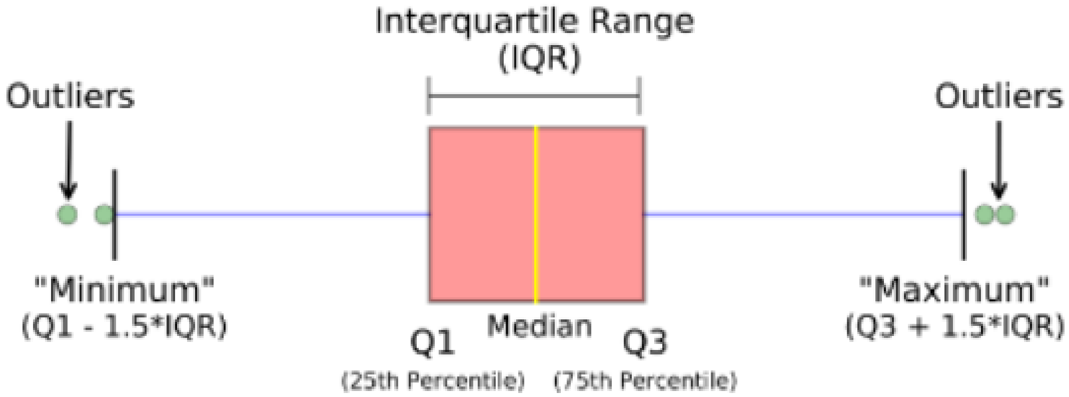
3. **Mode:**

- **Is the most repeated number in a group of numbers**
- **It can be set of numbers.**

4. **Med:**

- **the sum of the lower and upper limits of the class divided by two.**

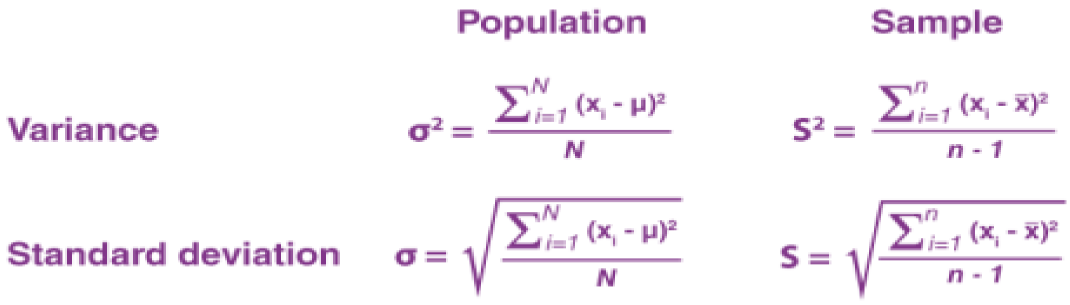
- **Histograms: The most common visual for quantitative data.**
- **There is also box plot: are useful for quickly comparing the spread of two data sets across some key metrics, like quartiles, maximum, and minimum.**



- **Measure of spread: One of the most common ways to measure the spread of data is the 5-Numbers-Summary**
    - **5-Numbers-Summary: gives values for calculating the range and interquartile range for a ordered dataset It consists of 5 values:**
        - **maximum: the biggest value in the data set**
        - **third quartile: the median of the values between the maximum and second quartile (75% of the data falls below it)**
        - **second quartile (median): the median of the values**
        - **first quartile : the median of the values between the minimum and second quartile (25% of the data falls below it)**
        - **minimum: the smallest value in the data set**
    - **The range is calculated: by subtracting the maximum from the minimum.**
    - **The interquartile range is calculated: by subtracting the values of the 3rd & 1st quartiles.**
    - **The spread of data is measured most commonly using a single value is with Standard deviation or with Variance.**

- Standard Deviation: How much each point on average varies from the mean of the points (EX: how much on average the distance of each of the employees of a company differs from the average distance all employees are from work). (IT IS THE SQRT OF VARIANCE)
- Variance: The average squared difference of each observation of data from the mean
- Calculating the standard deviation:
  - get the mean ($\bar{x}$)
  - square the difference between each value of the data set and the mean ($x_i - \bar{x}$)
  - get the average squared distance of each observation of the mean (variance)
  - square root the ending value and we get the standard deviation.
- The standard deviation is the sqrt of the variance.
- The higher the mean value is the lower the standard deviation and variance are.

  mu = mean xi = given numbers N = count of numbers



	Population	Sample
Variance	$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$	$S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}$
Standard deviation	$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$	$S = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}}$

- **Important Final Points:**

  - **The variance or the standard deviation is used to compare the spread of two different groups.**
  - **A set of data with higher variance is more spread out than a dataset with lower variance.**
  - **Be careful though, there might just be an outlier (or outliers) that is increasing the variance, when most of the data are actually very close.**
  - **When comparing the spread between two datasets, the units of each must be the same.**
  - **When data are related to money or the economy, higher variance (or standard deviation) is associated with higher risk.**
  - **The standard deviation is used more often in practice than the variance because it shares the units of the original dataset.**

- **Measures Of Shape: is how to use histograms to determine the shape associated with data.**
  - **here are 3 examples of histogram shapes:**
    - **Left skewed: the left most bin is smaller than the right most bin (Mean less than Median)**
    - **Right Skewed: the right most bin is smaller than the left most bin (Mean greater than Median)**
    - **Symmetrical : you can draw a line down the middle and have both sides mirroring (frequently normally distributed)**
    - **The mode of a distribution is essentially the tallest bar in a histogram.**

- **Outliers:**
  - Data points that fall very far from the rest of the data values in out dataset. the five number summary is better than the mean and standard deviation when outliers are present.
  - with outliers you should:
    - Note the impact they have on summary.
    - Remove / Fix them if they're typos.
    - Understand why they exist and their impact on questions we're trying to answer.
    - be careful when reporting and ask the right questions.

  - Identifying Outliers:
    - Sorting your values from low to high and checking minimum and maximum values.
    - Visualizing your data with a box plot and looking for outliers.
    - Using the interquartile range to create fences for your data.
  - An outlier must satisfy either of the following two conditions:
    - outlier < Q1 - 1.5(IQR)
    - outlier > Q3 + 1.5(IQR)
  - If we are dealing with a bell shaped data, we can learn a lot with the mean and the standard deviation in normally distributed data
  - Five number summary is better for skewed data.

- **Categorical: 1) nominal 2) ordinal**
- **Nominal:**

1. Nominal data simply names something without assigning it to an order in relation to other numbered objects or pieces of data.
2. An example of nominal data might be a "pass" or "fail" classification for each student's test result.

3. Nominal data provides some information about a group or set of events, even if that information is limited to mere counts.
4. **Nominal scales are also known as labels**

- **Ordinal:**

1. Ordinal data, unlike nominal data, involves some order; ordinal numbers stand in relation to each other in a ranked fashion.
2. For example, suppose you receive a survey from your favorite restaurant that asks you to provide.
3. feedback on the service you received. You can rank the quality of service as "1" for poor, "2" for below average, "3" for average, "4" for very good and "5" for excellent.
4. The data collected by this survey are examples of ordinal data. Here the numbers assigned have an order or rank; that is, a ranking of "4" is better than a ranking of "2."

- **Two types of data scientists collect are qualitative and quantitative.**
- **If a dataset is evenly distributed around the mean, this is certainly called a symmetric distribution.**

Notation:

- **Common math language used to communicate regardless of spoken language. (Essential to communicating ideas regarding data)**
- **Variables:**
  - **Random: Notated by a capital letter (They have many different values)**
  - **Observed: Notated by a lowercase letter with a subscript (signify a specific value)**

❖ Inferential Statistics:

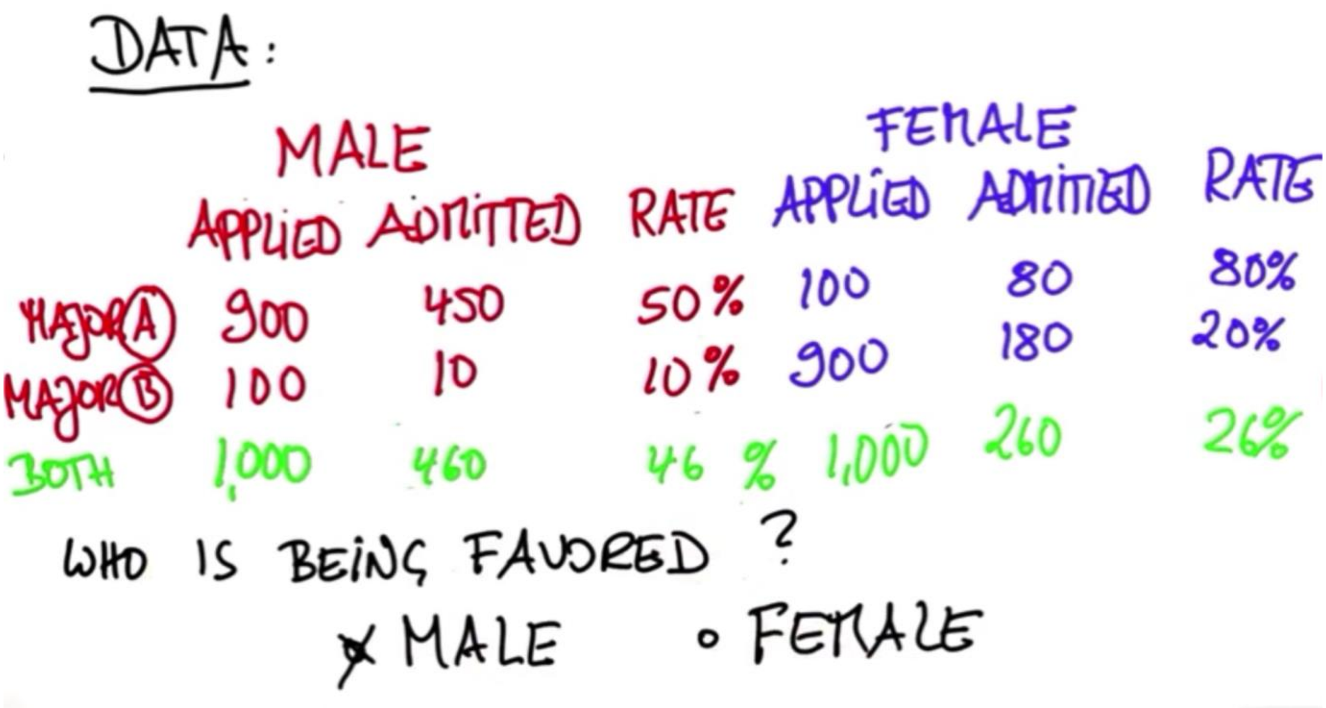
- Inferential statistics is a branch of statistics that makes the use of various analytical tools to draw inferences about the population data from sample data.
- Inferential statistics help to draw conclusions about the population Inferential statistics helps to develop a good understanding of the population data by analyzing the samples obtained from it.
- It helps in making generalizations about the population by using various analytical tests and tools. To pick out random samples that will represent the population accurately many sampling techniques are used.
- Inferential statistics can be classified into hypothesis testing and regression analysis.
- **It can be divided into:**
  - **Population** - our entire group of interest (A defined collection of individuals or objects about which we want to draw conclusions).
  - **Parameter** - numeric summary about a population (A numerical quantity measuring some aspect of a population).
  - **Sample** - subset of the population
  - **Statistic** - numeric summary about a sample
  - **Census** - The collection of information from the whole population.

Descriptive vs. Inferential Statistics:

- Descriptive statistics is about describing our collected data.
- Inferential statistics is about using our collected data to draw conclusions about a larger population.

Simpson’s paradox

1. Data Sometimes data can be tricky.
2. We shouldn’t judge the data from one view we should see all the possible views so that we can understand the data and give a fair decision about it.
3. To so such a thing we could use Simpson’s paradox method.



DATA:

	MALE			FEMALE		
	APPLIED	ADMITTED	RATE	APPLIED	ADMITTED	RATE
MAJOR(A)	900	450	50%	100	80	80%
MAJOR(B)	100	10	10%	900	180	20%
BOTH	1,000	460	46 %	1,000	260	26%

WHO IS BEING FAVORED ?

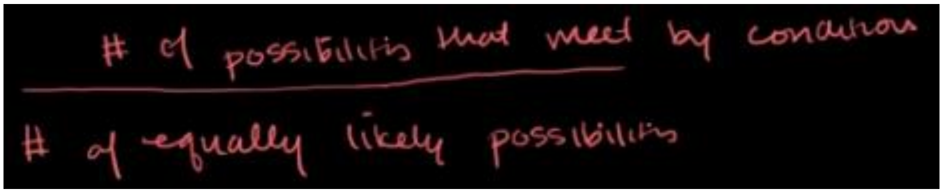
✗ MALE ○ FEMALE

4. As seen from this example although the rate of males is less than the females in every major, males are more favored than females by looking to the over total number.

probability

- It is the opposite of statistics as in statistics we use to analyze data, in probability we predict data using assumptions we make about it.
- P(A) == 1 – P(>A) → Not

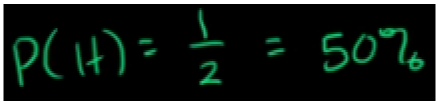
- **The best example for understanding probability is flipping a coin:**
  - There are two possible outcomes—heads or tails.
  - What's the probability of the coin landing on Heads? We can find out using the equation (*H*)=? P, left parenthesis, H, right parenthesis, equals, question mark. You might intuitively know that the likelihood is half/half, or 50%. But how do we work that out? Probability =



$$\frac{\text{\# of possibilities that meet my condition}}{\text{\# of equally likely possibilities}}$$

Formula for calculating the probability of certain outcomes for an event.

In this case:



$$P(H) = \frac{1}{2} = 50\%$$

Probability of a coin landing on heads

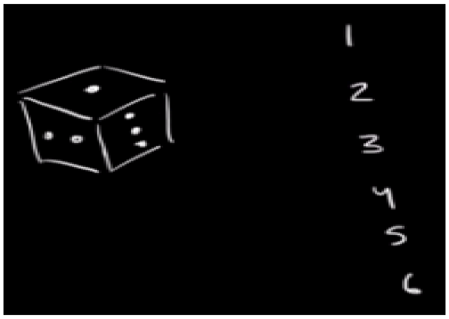
- **Probability of an event = (# of ways it can happen) / (total number of outcomes)**

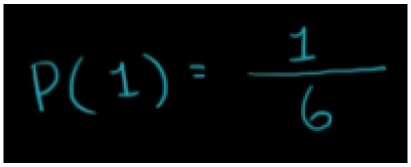
  P(A) = (# of ways A can happen) / (Total number of outcomes)

- **Example 1**

  There are six different outcomes.

  What's the probability of rolling a one?

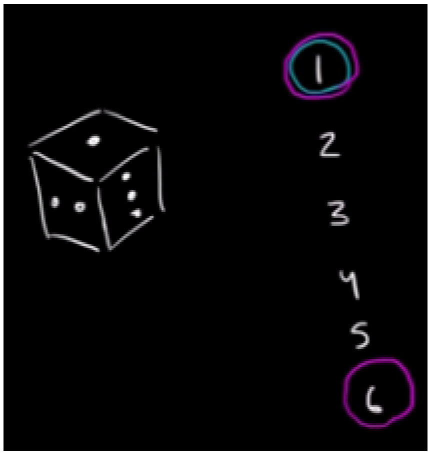




$$P(1) = \frac{1}{6}$$

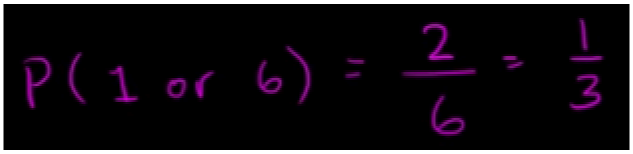
#Probability formula for rolling a '1' on a die

What’s the probability of rolling a one or a six?



#Probability of a 1 or a 6 outcome when rolling a die

Using the formula from above:



$$P(1 \text{ or } 6) = \frac{2}{6} = \frac{1}{3}$$

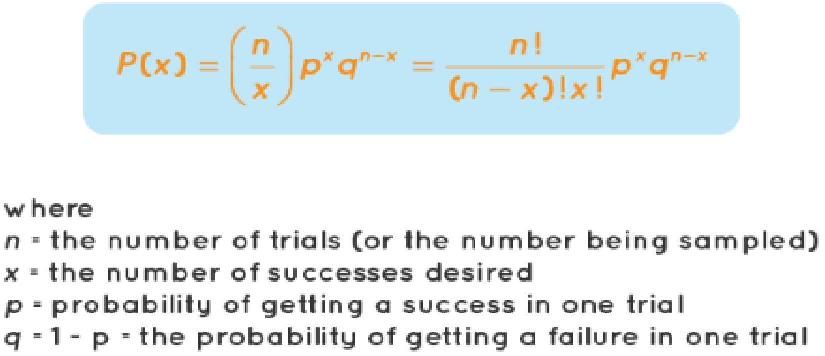
#Probability formula applied

What’s the probability of rolling an even number (i.e., rolling a two, four or a six)?

- To make it easier to calculate the possible probabilities we can use truth table.

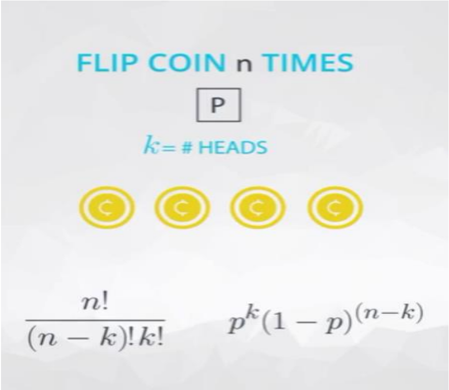
binomial distribution

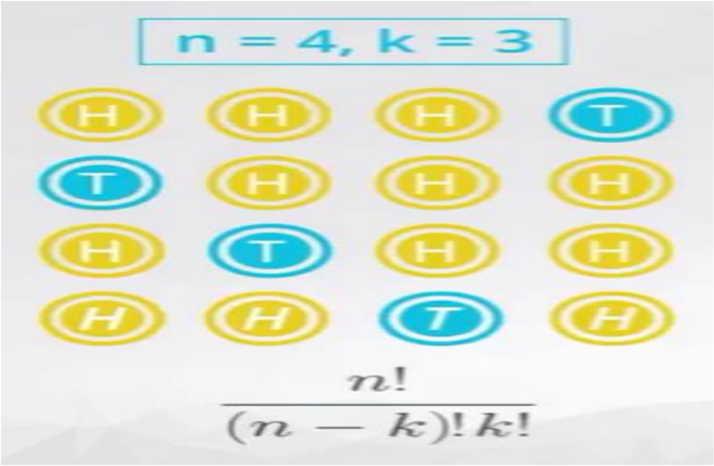
- In statistics and probability theory, the **binomial distribution** is the probability distribution that is discrete and applicable to events having only two possible results in an experiment, either success or failure. (The prefix “bi” means two, or twice). A few circumstances where we have binomial experiments are tossing a coin: head or tail, the result of a test: pass or fail, selected in an interview: yes/ no, or nature of the product: defective/non-defective. Such a distribution of a binomial random variable is called a binomial probability distribution.
- Binomial Distribution is a commonly used discrete distribution in statistics. The normal distribution as opposed to a binomial distribution is a continuous distribution. Let us learn the formula to calculate the Binomial distribution considering many experiments and a few solved examples for a better understanding.

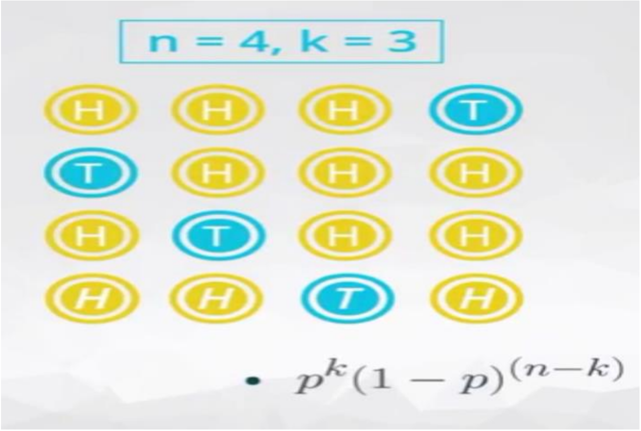


$$P(x) = \binom{n}{x} p^x q^{n-x} = \frac{n!}{(n-x)!x!} p^x q^{n-x}$$

where
n = the number of trials (or the number being sampled)
x = the number of successes desired
p = probability of getting a success in one trial
q = 1 - p = the probability of getting a failure in one trial


FLIP COIN n TIMES
P
k= # HEADS
C
C
C
C
$\frac{n!}{(n-k)!k!}$
$p^k(1-p)^{(n-k)}$


n = 4, k = 3
H H H T
T H H H
H T H H
H H T H
$\frac{n!}{(n-k)!k!}$


n = 4, k = 3
H H H T
T H H H
H T H H
H H T H
• $p^k(1-p)^{(n-k)}$