PROBABILITY AND STATISTICS

MODULE - 7

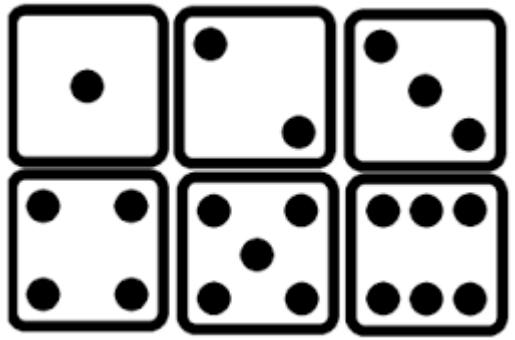
We will go through Probability and Statistics whereas they are key to understand, process and interpret the vast amount of data, we will deal with the basics of probability and statistics like Probability theory, Bayes theorem, distributions etc and their importance. Besides that, we will do hands-on with Numpy upon those concepts

WHAT IS THE MEANING OF DATA SCIENCE?

- Biologist - analyzing DNA sequences.
- Banker - predicting the stock market.
- Software engineer - programs and data structures
- machine learning scientist - models and algorithms.
- What is common ? - uncertainty.
- Therefore, data science is the subject of making decisions in uncertainty.
- The mathematics of analyzing uncertainty is probability. It is the tool to help us model, analyze, and predict random events.

WHAT IS PROBABILITY?

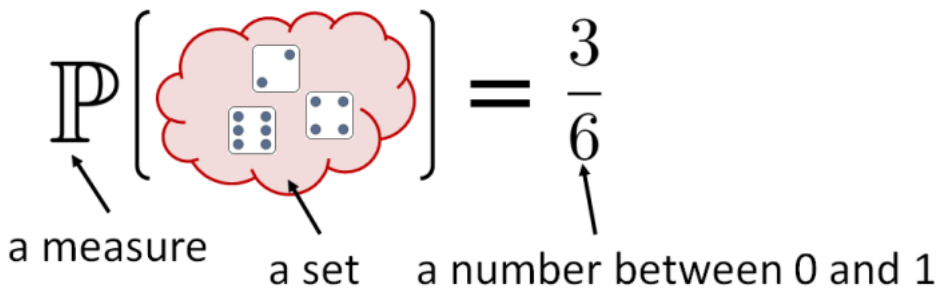
- Data and probability are inseparable.
  - Data is the computational side of the story, whereas probability is the theoretical side of the story.
- Any data science practice must be built on the foundation of probability.
- Probability can also be stated as the relative frequency of an outcome.
  - For example, flipping a fair coin has a 1/2 probability of getting a head because if you flip the coin infinitely many times, you will have half of the time getting a head.
- Probability of throwing a die.
  - What is the probability of getting dice 3.
  - What is the probability that you get a number that is "less than 5"?
  - What is the probability that you get a number that is "less than 5" and is "an even number"



- What is the sample space in this problem?
- A sample space is the set containing all possible outcomes

COMPONENTS OF PROBABILITY

- There is a sample space, which is the set that contains all the possible outcomes.
- There is an event, which is a subset inside the sample space.
- Two events E1 and E2 can be combined to construct another event E that is still a subset inside the sample space.
- Probability is a number assigned by certain rules such that it describes the relative size of the event E compared with the sample space
- Probability is a measure of the size of a set. Whenever we talk about probability, it has to be the probability of a set.



PROBABILITY SPACE

- Sample Space: The set of all possible outcomes from an experiment.
- Event Space F: The collection of all possible events. An event E is a subset in sample space that defines an outcome or a combination of outcomes.
- Probability Law P: A mapping from an event E to a number P[E] which, ideally, measures the size of the event.

SAMPLE SPACE

Coin flip: $\Omega = \{H,\ T\}$.

Throw a die: $\Omega = \{\boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot\}$.

Paper / scissor / stone: $\Omega = \{\text{paper}, \text{scissor}, \text{stone}\}$.

Draw an even integer: $\Omega = \{2, 4, 6, 8, \ldots\}$.

If the experiment contains flipping a coin and throwing a die, then the sample space is

$$\Big\{(H,\boxdot),(H,\boxdot),(H,\boxdot),(H,\boxdot),(H,\boxdot),(H,\boxdot),$$
$$(T,\boxdot),(T,\boxdot),(T,\boxdot),(T,\boxdot),(T,\boxdot),(T,\boxdot)\Big\}.$$

In this sample space, each element is a pair of outcomes.

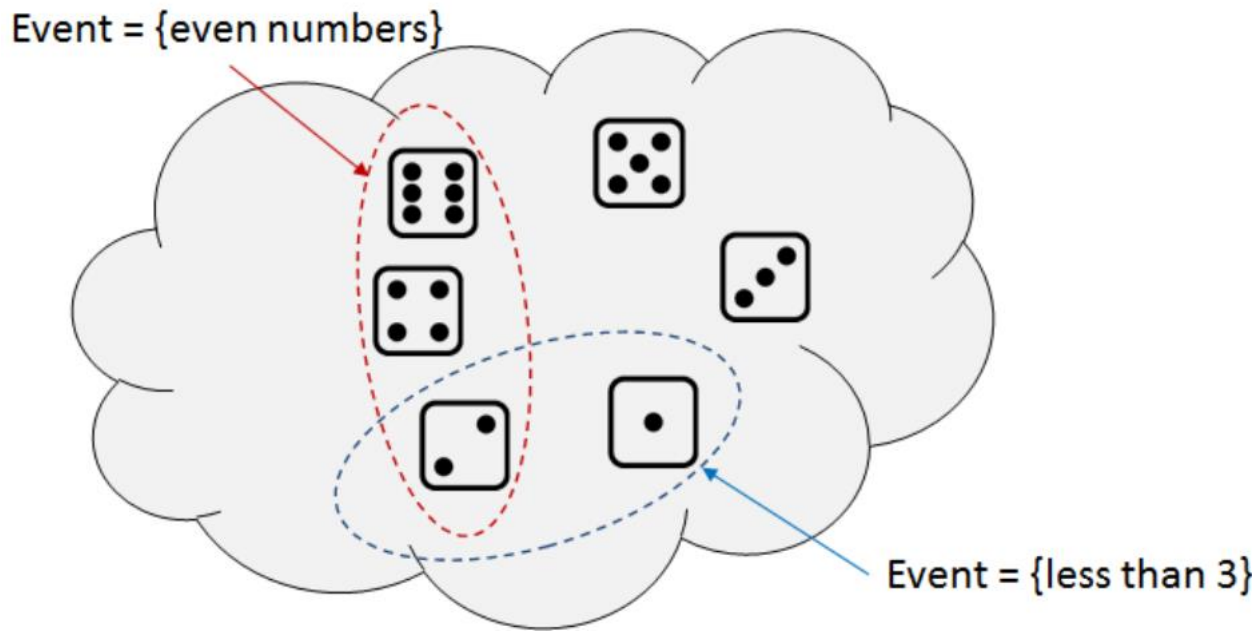
EVENT SPACE

- An event E is a subset in the sample space. The set of all possible events is called Event Space

Throw a die. Let $\Omega = \{\boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot\}$

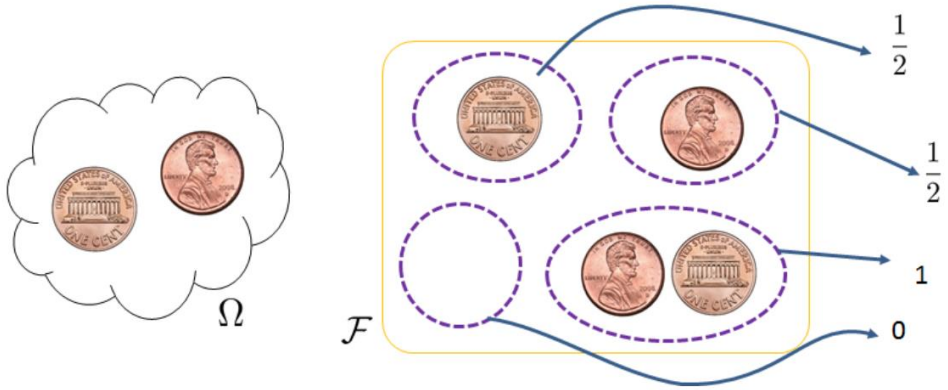
$E_1 = \{\text{even numbers}\} = \{\boxdot, \boxdot, \boxdot\}.$

$E_2 = \{\text{less than } 3\} = \{\boxdot, \boxdot\}$



PROBABILITY LAW

- The probability law is a function.
- Its job is to assign a number to an event.
- The input to P is an event E, which is a subset in Sample Space and an element in Event Space.
- The output of P is a number between 0 and 1, which we call the probability.
- For a probability to be valid, it must satisfy the axioms of probability.



Consider flipping a coin. The event space is $\mathcal{F} = \{\emptyset, \{H\}, \{T\}, \Omega\}$. We can define the probability law as

$$\mathbb{P}[\emptyset] = 0, \quad \mathbb{P}[\{H\}] = \frac{1}{2}, \quad \mathbb{P}[\{T\}] = \frac{1}{2}, \quad \mathbb{P}[\Omega] = 1,$$

AXIOMS OF PROBABILITY

- The necessary restrictions on assigning a probability to an event are collectively known as the axioms of probability

A **probability law** *is a function* $\mathbb{P} : \mathcal{F} \to [0, 1]$ *that maps an event* A *to a real number in* $[0, 1]$. *The function must satisfy the* **axioms of probability**:

I. **Non-negativity**: $\mathbb{P}[A] \geq 0$, *for any* $A \subseteq \Omega$.

II. **Normalization**: $\mathbb{P}[\Omega] = 1$.

III. **Additivity**: *For any disjoint sets* $\{A_1, A_2, \ldots\}$, *it must be true that*

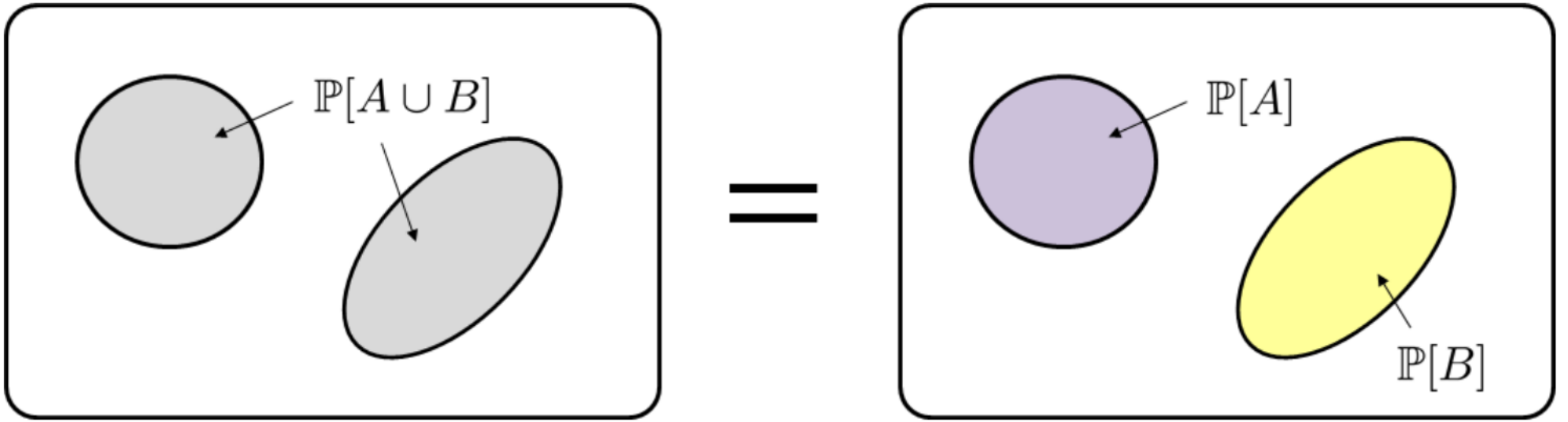
$$\mathbb{P}\left[\bigcup_{i=1}^{\infty} A_i\right] = \sum_{i=1}^{\infty} \mathbb{P}[A_i].$$

Why these three axioms?

- Axiom I (Non-negativity) ensures that probability is never negative.
- Axiom II (Normalization) ensures that probability is never greater than 1.
- Axiom III (Additivity) allows us to add probabilities when two events do not overlap.

The probability of getting {⚄, ⚅} is

$$\mathbb{P}[\{⚄, ⚅\}] = \mathbb{P}[\{⚄\} \cup \{⚅\}] = \mathbb{P}[\{⚄\}] + \mathbb{P}[\{⚅\}] = \frac{1}{6} + \frac{1}{6} = \frac{2}{6}$$

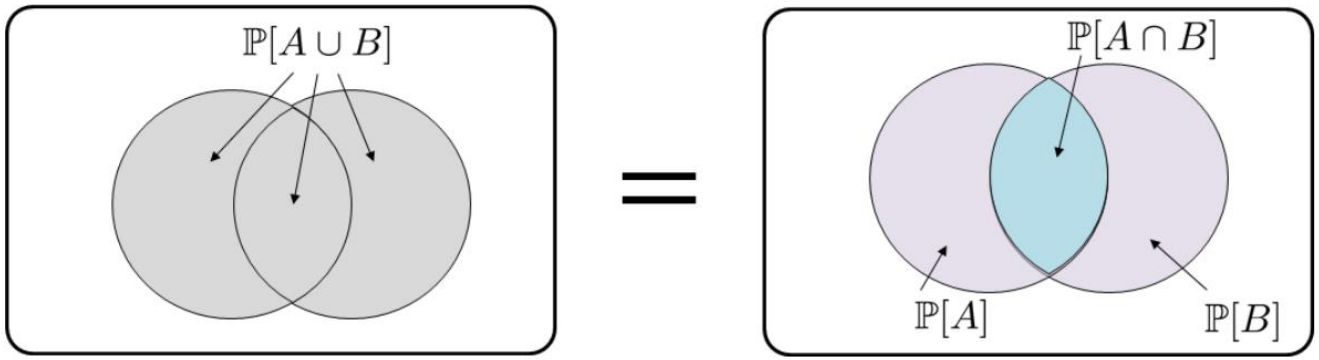


Axiom III says $\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B]$ if $A \cap B = \emptyset$.

Let $\Omega = \{\boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot\}$ be the sample space of a fair die. Let $A = \{\boxdot, \boxdot, \boxdot\}$ and $B = \{\boxdot, \boxdot, \boxdot\}$. Then

$$\mathbb{P}[A \cup B] = \mathbb{P}[\{\boxdot, \boxdot, \boxdot, \boxdot, \boxdot\}] = \frac{5}{6}.$$

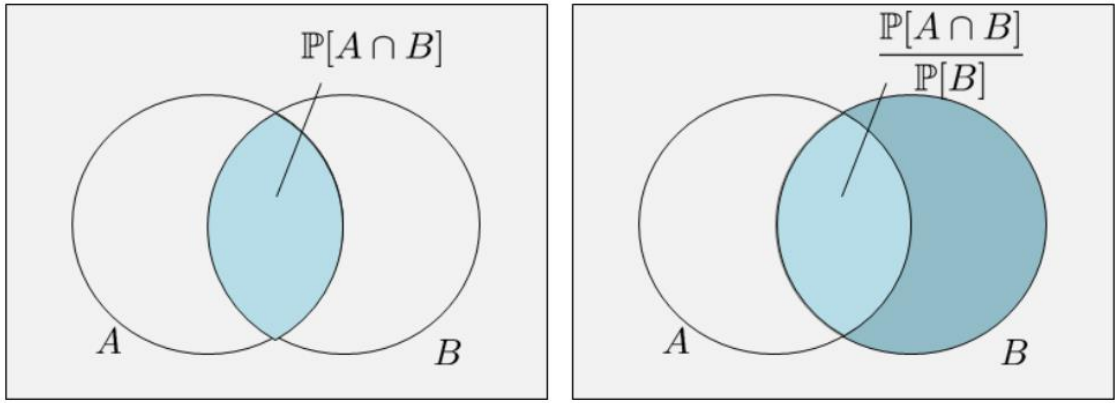
$$\begin{aligned}\mathbb{P}[A \cup B] &= \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B] \\ &= \mathbb{P}[\{\boxdot, \boxdot, \boxdot\}] + \mathbb{P}[\{\boxdot, \boxdot, \boxdot\}] - \mathbb{P}[\{\boxdot\}] \\ &= \frac{3}{6} + \frac{3}{6} - \frac{1}{6} = \frac{5}{6}.\end{aligned}$$



CONDITIONAL PROBABILITY

- In Data Science, we are interested in the relationship between two or more events.
- For example, an event A may cause B to happen, and B may cause C to happen.
  - A legitimate question in probability is then: If A has happened, what is the probability that B also happens?
- the conditional probability of A given B is the ratio of P[A ∩ B] to P[B]. It is the probability that A happens when we know that B has already happened. Since B has already happened, the event that A has also happened is represented by A∩B.

$$\mathbb{P}[A \mid B] \stackrel{\text{def}}{=} \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}$$



Consider throwing a die. Let

$$A = \{\text{getting a 3}\} \quad \text{and} \quad B = \{\text{getting an odd number}\}.$$

Find $\mathbb{P}[A \mid B]$ and $\mathbb{P}[B \mid A]$.

Solution. The following probabilities are easy to calculate:

$$\mathbb{P}[A] = \mathbb{P}[\{\boxdot\}] = \frac{1}{6}, \qquad \text{and} \qquad \mathbb{P}[B] = \mathbb{P}[\{\boxdot, \boxdot, \boxdot\}] = \frac{3}{6}.$$

Also, the intersection is

$$\mathbb{P}[A \cap B] = \mathbb{P}[\{\boxdot\}] = \frac{1}{6}.$$

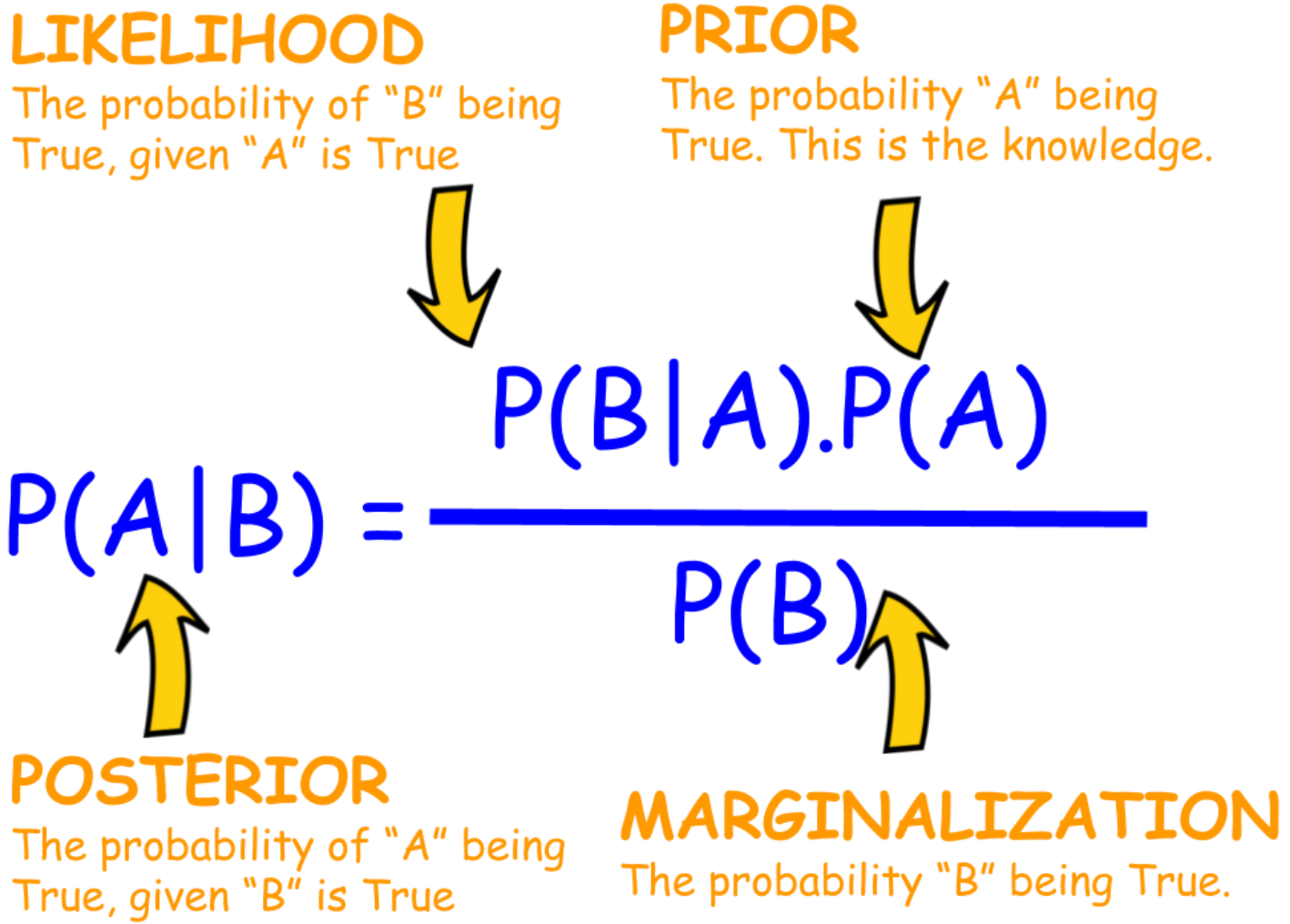
Given these values, the conditional probability of A given B can be calculated as

$$\mathbb{P}[A \mid B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]} = \frac{\frac{1}{6}}{\frac{3}{6}} = \frac{1}{3}.$$

$$\mathbb{P}[B \mid A] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[A]} = 1.$$

Therefore, if we know that we have rolled a 3, then the probability for this number being an odd number is 1.

BAYES THEOREM



BAYES THEOREM

Theorem (**Bayes' theorem**). *For any two events A and B such that $\mathbb{P}[A] > 0$ and $\mathbb{P}[B] > 0$,*

$$\mathbb{P}[A \mid B] = \frac{\mathbb{P}[B \mid A]\,\mathbb{P}[A]}{\mathbb{P}[B]}.$$

Proof. By the definition of conditional probabilities, we have

$$\mathbb{P}[A \mid B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]} \quad \text{and} \quad \mathbb{P}[B \mid A] = \frac{\mathbb{P}[B \cap A]}{\mathbb{P}[A]}.$$

Rearranging the terms yields

$$\mathbb{P}[A \mid B]\mathbb{P}[B] = \mathbb{P}[B \mid A]\mathbb{P}[A],$$

which gives the desired result by dividing both sides by $\mathbb{P}[B]$.

$$P(\text{Cancer}|\text{Symptoms}) = \frac{P(\text{Symptoms}|\text{Cancer})P(\text{Cancer})}{P(\text{Symptoms})}$$

$$= \frac{P(\text{Symptoms}|\text{Cancer})P(\text{Cancer})}{P(\text{Symptoms}|\text{Cancer})P(\text{Cancer}) + P(\text{Symptoms}|\text{Non-Cancer})P(\text{Non-Cancer})}$$

$$= \frac{1 \times 0.00001}{1 \times 0.00001 + (10/99999) \times 0.99999} = \frac{1}{11} \approx 9.1\%$$

probability a hypothesis is true given the evidence

probability a hypothesis is true (before any evidence is present)

probability of seeing the evidence if the hypothesis is true

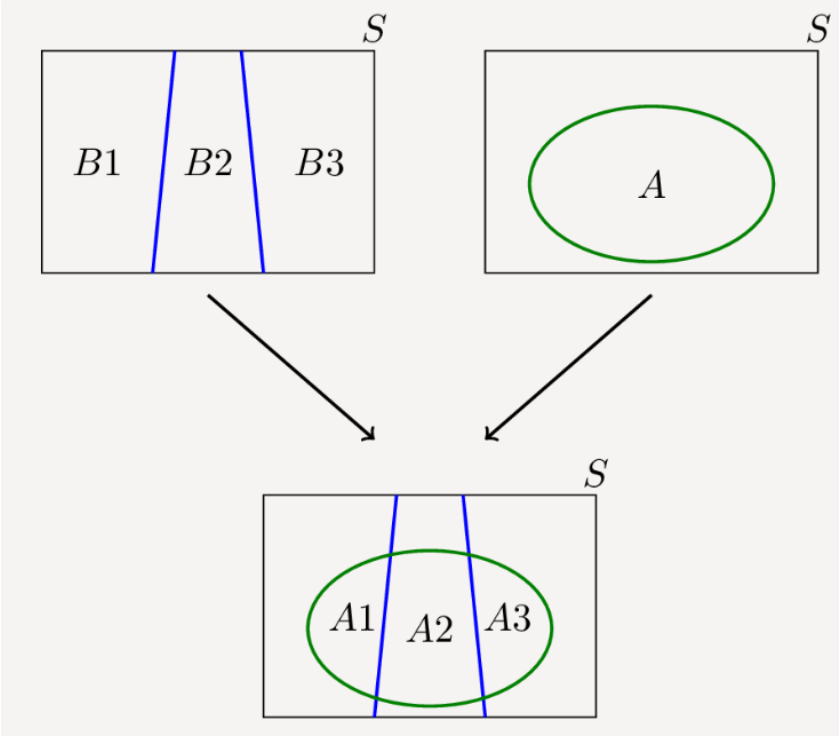
probability of observing the evidence

$$P(H/E) = \frac{P(H)\,P(E/H)}{P(E)}$$

LAW OF TOTAL PROBABILITY

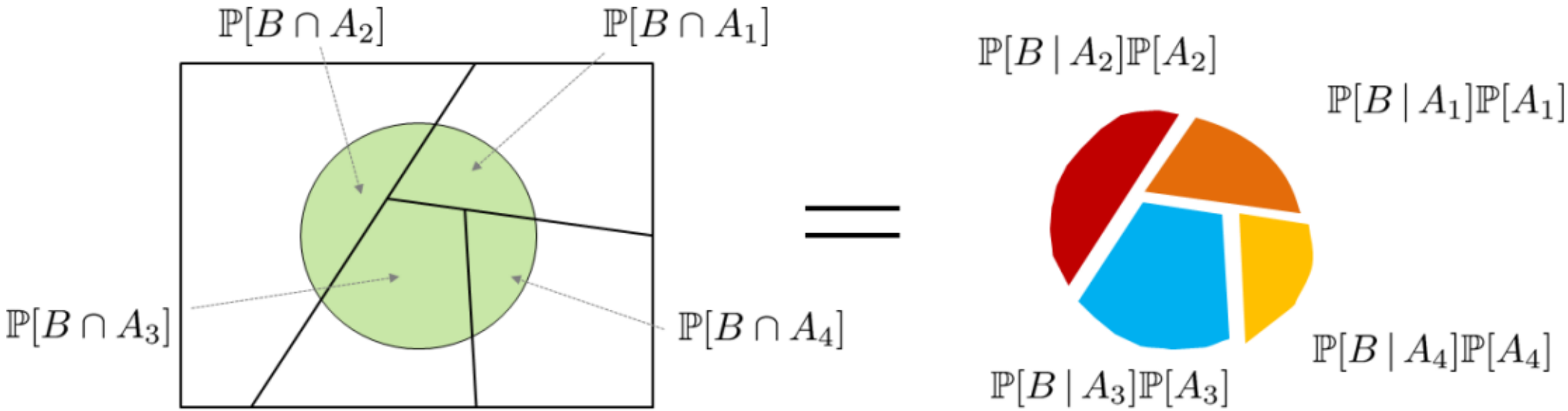
If $B_1, B_2, B_3, \cdots$ is a partition of the sample space S, then for any event A we have

$$P(A) = \sum_i P(A \cap B_i) = \sum_i P(A|B_i)P(B_i).$$



$$A_1 = A \cap B_1,$$
$$A_2 = A \cap B_2,$$
$$A_3 = A \cap B_3.$$

$$P(A) = P(A_1) + P(A_2) + P(A_3)$$



The law of total probability decomposes the probability $\mathbb{P}[B]$ into multiple conditional probabilities $\mathbb{P}[B \mid A_i]$. The probability of obtaining each $\mathbb{P}[B \mid A_i]$ is $\mathbb{P}[A_i]$.

EXAMPLE

I have three bags that each contain 100 marbles:

- Bag 1 has 75 red and 25 blue marbles;
- Bag 2 has 60 red and 40 blue marbles;
- Bag 3 has 45 red and 55 blue marbles.

I choose one of the bags at random and then pick a marble from the chosen bag, also at random. What is the probability that the chosen marble is red?

Let R be the event that the chosen marble is red. Let B_i be the event that I choose Bag i. We already know that

$$P(R|B_1) = 0.75,$$

$$P(R|B_2) = 0.60,$$

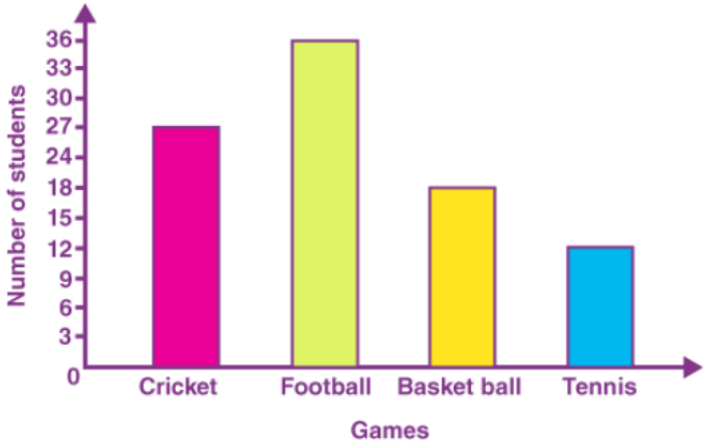
$$P(R|B_3) = 0.45$$

We choose our partition as B_1, B_2, B_3. Note that this is a valid partition because, firstly, the B_i's are disjoint (only one of them can happen), and secondly, because their union is the entire sample space as one the bags will be chosen for sure, i.e., $P(B_1 \cup B_2 \cup B_3) = 1$. Using the law of total probability, we can write

$$\begin{aligned} P(R) &= P(R|B_1)P(B_1) + P(R|B_2)P(B_2) + P(R|B_3)P(B_3) \\ &= (0.75)\tfrac{1}{3} + (0.60)\tfrac{1}{3} + (0.45)\tfrac{1}{3} \\ &= 0.60 \end{aligned}$$

WHAT IS A HISTOGRAM?

- In statistics, a histogram is a graphical representation of the distribution of data.
- It is represented by a set of rectangles, adjacent to each other, where each bar represent a kind of data.
- When numerals are repeated in statistical data, this repetition is known as Frequency and which can be written in the form of a table, called a frequency distribution.
- A Frequency distribution can be shown graphically by using different types of graphs and a Histogram is one among them.



RANDOM VARIABLES, PMF AND CDF

- When working on a data analysis problem, one of the biggest challenges is the disparity between the theoretical tools we learn in school and the actual data our boss hands to us.

Key Concept 1: What are random variables?

Random variables are mappings from events to numbers.

Key Concept 2: What are probability mass functions (PMFs)?

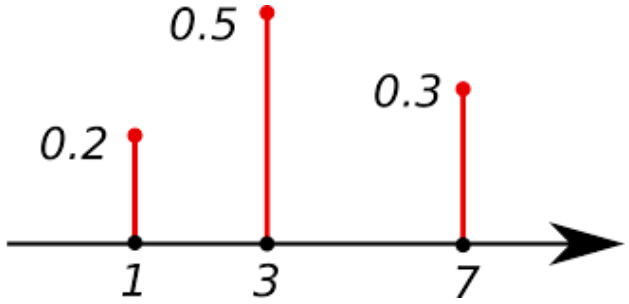
Probability mass functions are the ideal histograms of random variables.

Key Concept 3: What is expectation?

Expectation = Mean = Average computed from a PMF.

RANDOM VARIABLES, PMF AND CDF

- The sample space and event space are all based on statements.
  - eg: getting a head when flipping a coin. or winning the game
- How to convert these statements to numbers?
- **Random Variables** - are mappings from events to numbers.
- Next task is to assign a probability to the RV to perform future computations.
  - This is done using PMF.



- **Probability Mass Functions** are ideal histograms of random variables.
- A histogram has two axes: The x-axis denotes the set of states and the y-axis denotes the probability. For each of the states that the random variable possesses, the histogram tells us the probability of getting a particular state. The PMF is the ideal histogram of a random variable. It provides a complete characterization of the random variable.
- Information can be pulled out from PMF

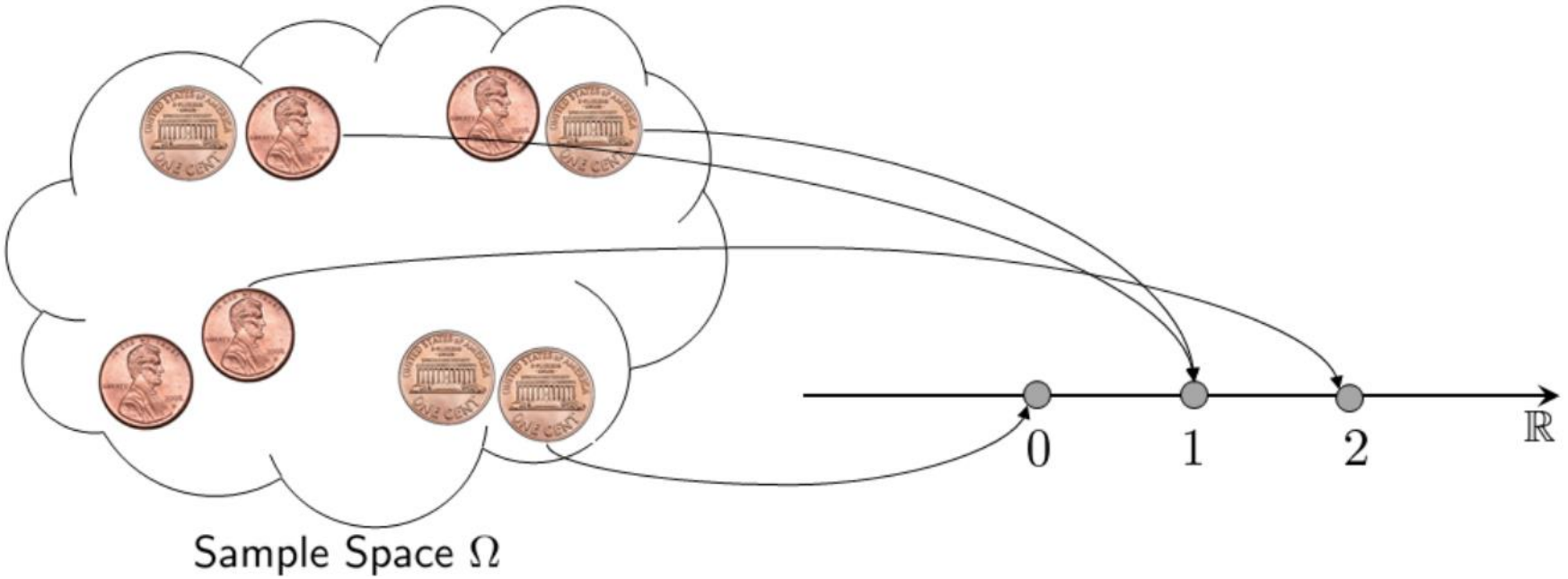
A MOTIVATING EXAMPLE FOR RANDOM VARIABLE

Consider an experiment with 4 outcomes $\Omega = \{\clubsuit, \diamondsuit, \heartsuit, \spadesuit\}$. We want to construct the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The sample space Ω is already defined. The event space $\mathcal{F}$ is the set of all possible subsets in Ω, which, in our case, is a set of 2^4 subsets. For the probability law $\mathbb{P}$, let us assume that the probability of obtaining each outcome is

$$\mathbb{P}[\{\clubsuit\}] = \frac{1}{6}, \quad \mathbb{P}[\{\diamondsuit\}] = \frac{2}{6}, \quad \mathbb{P}[\{\heartsuit\}] = \frac{2}{6}, \quad \mathbb{P}[\{\spadesuit\}] = \frac{1}{6}.$$

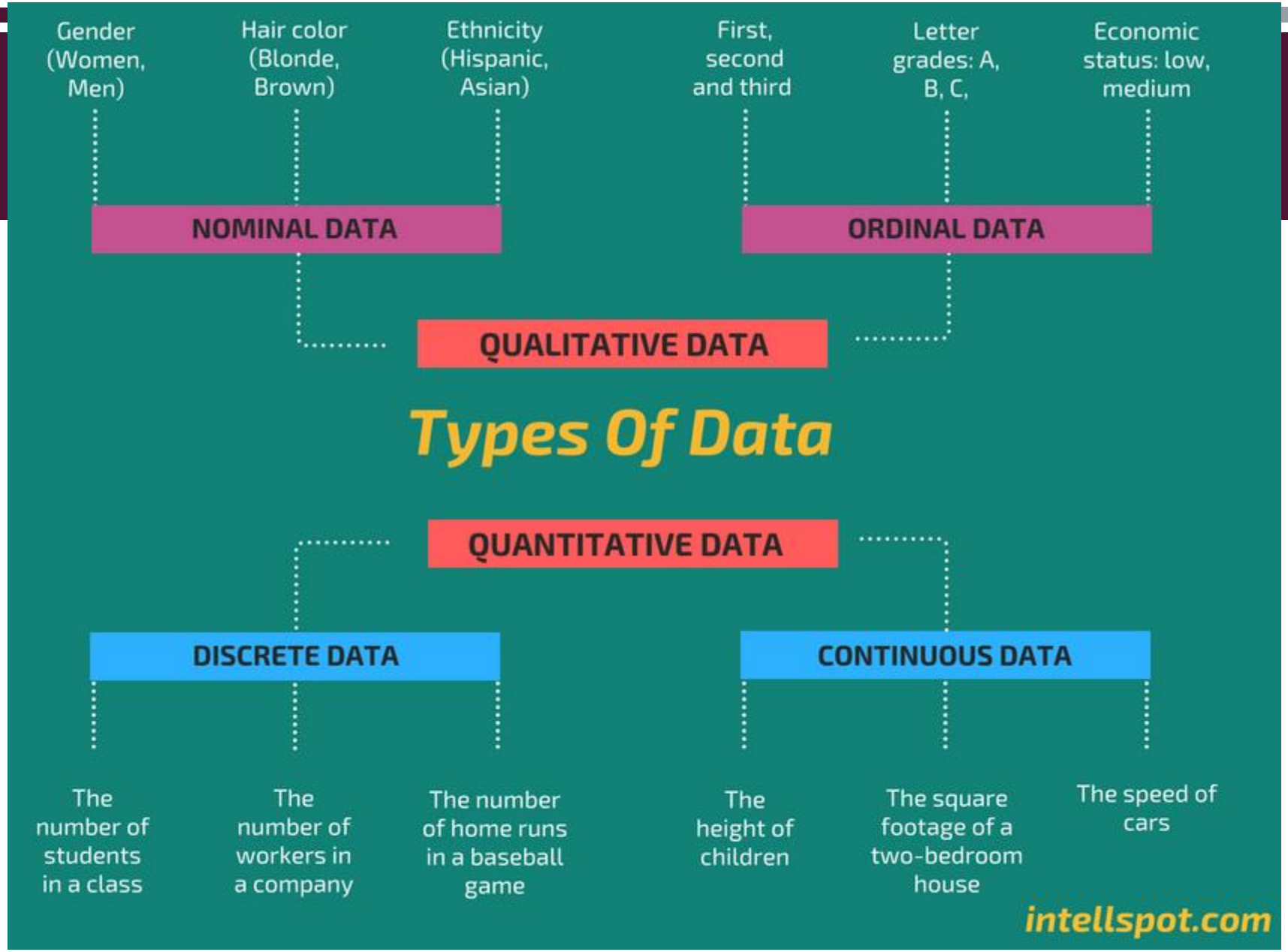
$$\mathbb{P}[X = 1] = \frac{1}{6}, \quad \mathbb{P}[X = 2] = \frac{2}{6}, \quad \mathbb{P}[X = 3] = \frac{2}{6}, \quad \mathbb{P}[X = 4] = \frac{1}{6}.$$

- A random variable X is a function mapping from a sample space to real number. (maps an outcome to a number)
- X is a function, but it is referred as a variable. X is a variable because X has multiple states.


0
1
2
ℝ
Sample Space Ω

TYPES OF RANDOM VARIABLES

- Discrete
- Continuous


Gender (Women, Men)
Hair color (Blonde, Brown)
Ethnicity (Hispanic, Asian)
First, second and third
Letter grades: A, B, C,
Economic status: low, medium
NOMINAL DATA
ORDINAL DATA
QUALITATIVE DATA
Types Of Data
QUANTITATIVE DATA
DISCRETE DATA
CONTINUOUS DATA
The number of students in a class
The number of workers in a company
The number of home runs in a baseball game
The height of children
The square footage of a two-bedroom house
The speed of cars
intellspot.com

Data description
Gender: Male/Female
T-shirt sizes: small, medium, large
Languages of the world: English, Hindi, Telugu, etc
Education Background, Elementary, High school, undergraduate, graduate, post graduate
Number of students in class
Heights of students in a class
Temperature
Atmospheric Pressure
Speed of a Vehicle
Weight
Customer Satisfaction

Data description	Data type
Gender: Male/Female	Qualitative - Nominal
T-shirt sizes: small, medium, large	Qualitative - Ordinal
Languages of the world: English, Hindi, Telugu, etc	Qualitative – Nominal
Education Background, Elementary, High school, undergraduate, graduate, post graduate	Qualitative – Ordinal
Number of students in class	Quantitative – Discrete
Heights of students in a class	Quantitative – Continuous
Temperature	Quantitative – Continuous
Atmospheric Pressure	Quantitative – Continuous
Speed of a Vehicle	Quantitative – Continuous
Weight	Quantitative – Continuous
Customer Satisfaction	Qualitative - Ordinal

PROBABILITY MASS FUNCTION

- Random variables are mappings that translate events to numbers. After the translation, we have a set of numbers denoting the states of the random variables.
- Each state has a different probability of occurring.
- The probabilities are summarized by a function known as the probability mass function (PMF).
- Sum of all the PMF values should be equal to 1

Flip a coin twice. The sample space is $\Omega = \{\text{HH, HT, TH, TT}\}$. We can assign a random variable X = number of heads. Therefore,

$$X(\text{“HH”}) = 2, X(\text{“TH”}) = 1, X(\text{“HT”}) = 1, X(\text{“TT”}) = 0.$$

So the random variable X takes three states: 0, 1, 2. The PMF is therefore

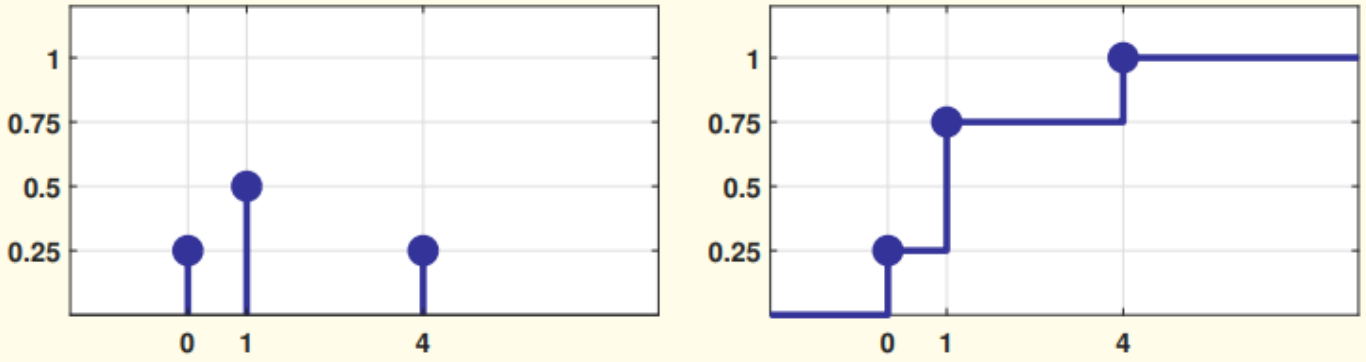
$$p_X(0) = \mathbb{P}[X = 0] = \mathbb{P}[\{\text{“TT”}\}] = \frac{1}{4},$$
$$p_X(1) = \mathbb{P}[X = 1] = \mathbb{P}[\{\text{“TH”}, \text{“HT”}\}] = \frac{1}{2},$$
$$p_X(2) = \mathbb{P}[X = 2] = \mathbb{P}[\{\text{“HH”}\}] = \frac{1}{4}.$$

CUMULATIVE DISTRIBUTIVE FUNCTIONS

Consider a random variable X with PMF $p_X(0) = \frac{1}{4}$, $p_X(1) = \frac{1}{2}$ and $p_X(4) = \frac{1}{4}$. The CDF of X can be computed as

$$F_X(0) = \mathbb{P}[X \leq 0] = p_X(0) = \frac{1}{4},$$
$$F_X(1) = \mathbb{P}[X \leq 1] = p_X(0) + p_X(1) = \frac{3}{4},$$
$$F_X(4) = \mathbb{P}[X \leq 4] = p_X(0) + p_X(1) + p_X(4) = 1.$$

As shown in **Figure 3.13**, the CDF of a discrete random variable is a staircase function.



(a) PMF $p_X(k)$ (b) CDF $F_X(k)$

Illustration of a PMF and a CDF.

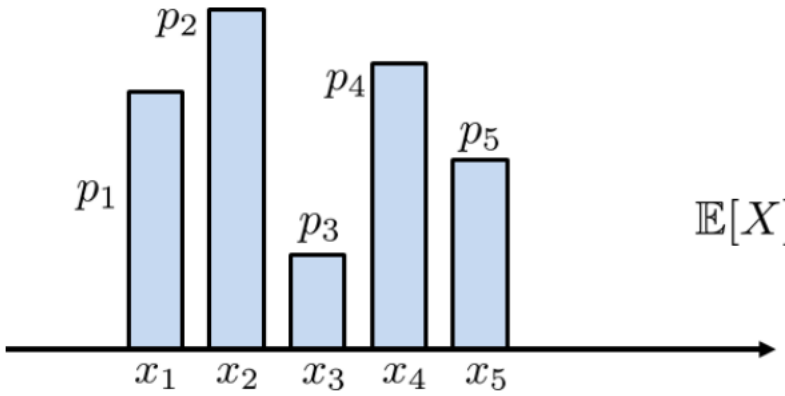
EXPECTATION

- When analyzing data, it is often useful to extract certain key parameters such as the mean and the standard deviation. The mean and the standard deviation can be seen from the lens of random variables.

The **expectation** *of a random variable X is*

$$\mathbb{E}[X] = \sum_{x \in X(\Omega)} x\, p_X(x).$$

$$\mathbb{E}[X] = \underbrace{\sum_{x \in X(\Omega)}}_{\text{sum over all states}} \underbrace{x}_{\text{a state } X \text{ takes}} \underbrace{p_X(x)}_{\text{the percentage}}$$

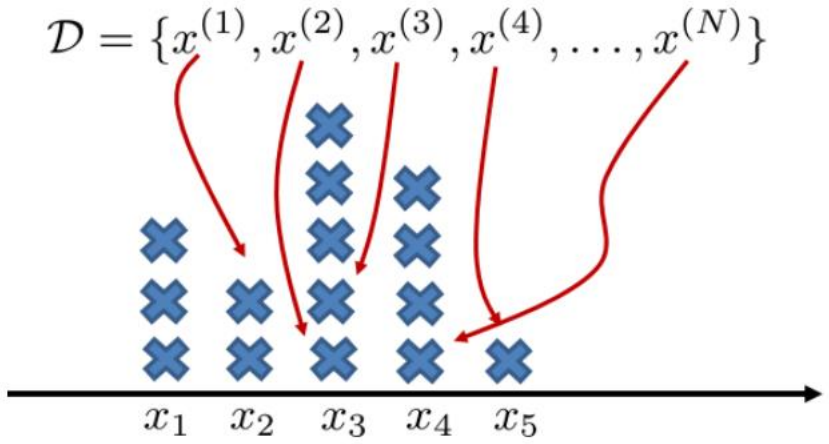


$$\mathbb{E}[X] = p_1 x_1 + \ldots + p_5 x_5$$

the average of a dataset $\mathcal{D} = \{x^{(1)}, x^{(2)}, \ldots, x^{(N)}$ average $= \frac{1}{N} \sum_{n=1}^{N} x^{(n)}$

in a typical dataset, these N samples often take distinct values.

average $= \frac{1}{N} \sum_{k=1}^{K}$ value $x_k \quad \times \quad$ number of samples with value x_k



WHY STATISTICS?

To summarize data to shape how we make decisions.

> Decisions about what and why? We usually want to decide on certain features that we can vary in order to predict, or drive, certain outcome. For example, Do costume affect performance? Or Is tumor size a good estimator of malignancy? etc.

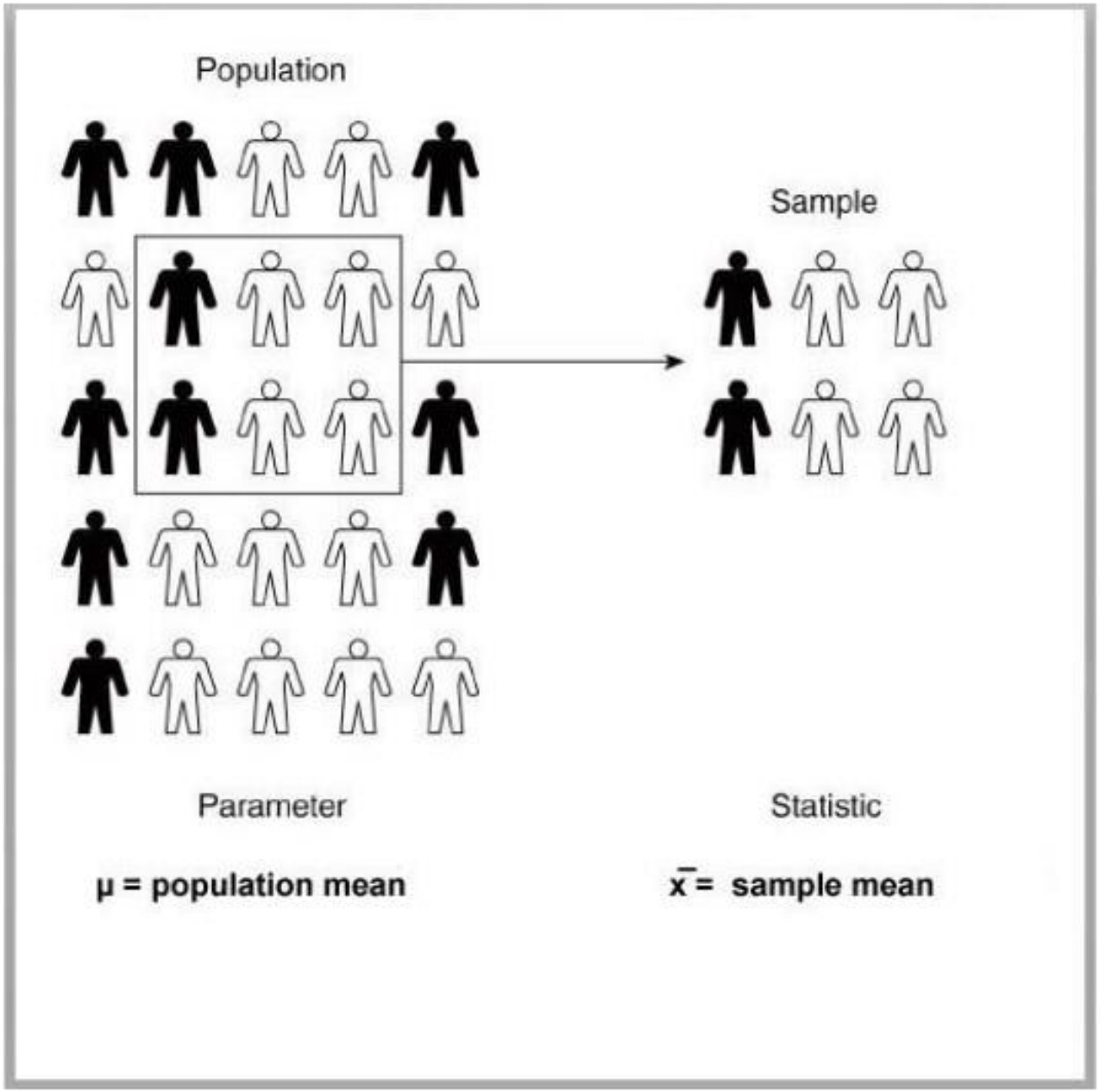
Can we measure everything? What about pain, happiness, satisfaction, performance, intelligence etc.?

These are all abstract concepts that do not have a measurable definition. However, in trying to understand it we need to give an OPERATIONAL definition to it which we call a **construct**.

For example, we can measure intelligence by interpreting scores on IQ test. We can review quality of product by star ratings given by its users and so on.

The operational definition rely upon some realistic HYPOTHESIS. It is a set of assumptions we make based on domain knowledge.

POPULATION VS SAMPLE



SAMPLING

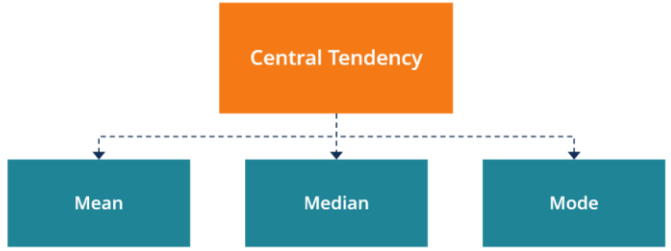
Process of selecting **RANDOMLY** small portion of the population and collecting observations regarding to those.

If we do not do it randomly we will be **BIASED**.

We want to describe our population. And in Statistical terms we do so by measuring the **central tendency** and **variability** of our observations.

When measured for a population, we call then Population **Parameter**. But, when done for a sample, we call them Sample **Statistic**

MEASURE OF CENTRAL TENDENCY



- A measure of central tendency is a single value that attempts to describe data by identifying the central position within that data.

Mean / Average: Population mean μ is defined as an average over all values. If the data is for a sample, we call it $\bar{x}$. Mean always only apply to numerical data.

$$\mu = \frac{1}{n}\sum_{i} x_i$$

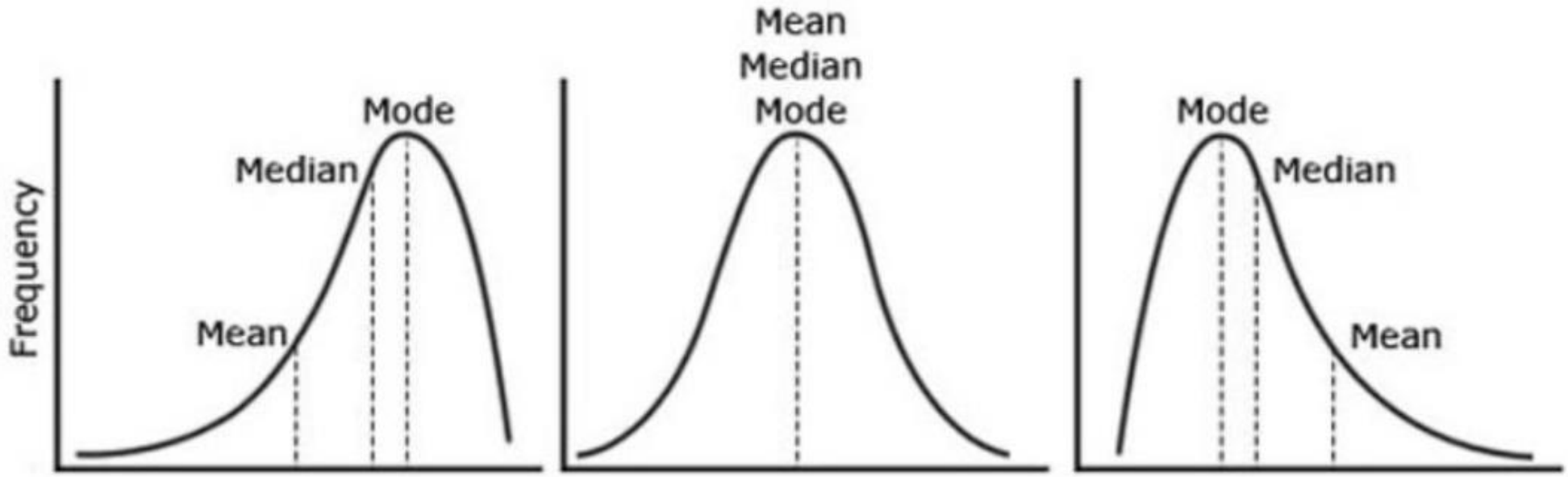
Where x_i is the i^{th} data value, and n is the count of data items.

Median

It refers to the middle value of data in sorted order. If count of data items is even it is equal to the average of middle two values. It applies to both numeric and categorical-ordinal data.

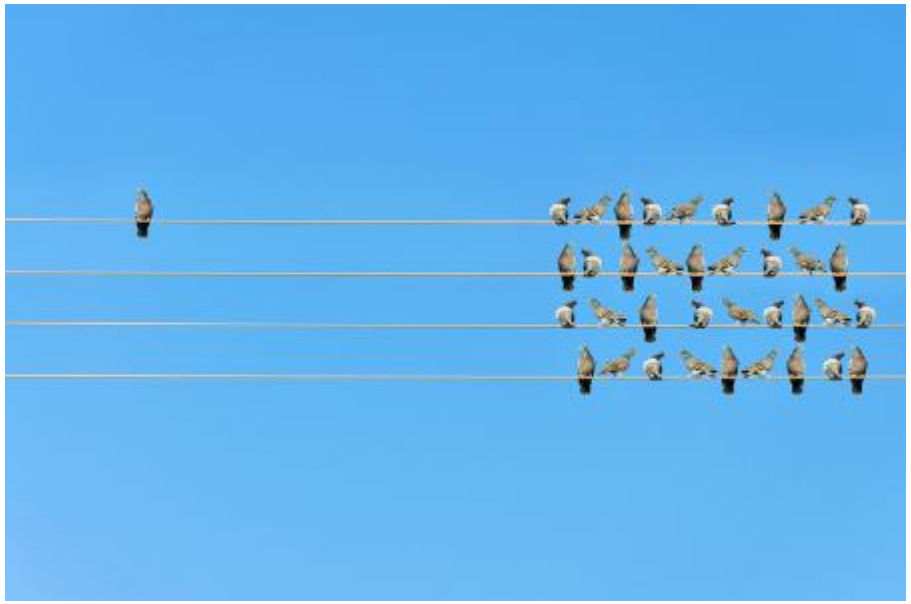
Mode

It refers to the data that occurs most frequently in the dataset. It applies to all types of data.


Frequency
Mode
Median
Mean
Mean
Median
Mode
Mode
Median
Mean

OUTLIERS

> Average income per month of people in city X is Rs. 50K. However 90% of the population is below poverty line. Is this possible?



VARIABILITY

Just a measure of central tendency is not sufficient. Consider:

- 1, 1, 1, 1, 1, 1, 1, 1, 1, 1
- -4, -3, -2, -1, 1, 1, 3, 4, 5, 6

Both of them have mean and median and mode of 1. However the first one has no variation in the data (all values are 1) where as the second one has a lot of variation.

Variance can be calculated only for numeric data. For non numeric data one may have to give some numeric values to each of the categorical label.

E.g. T-Shirt sizes can be 1,2,3,4,5 instead on XS,S,M,L,XL.

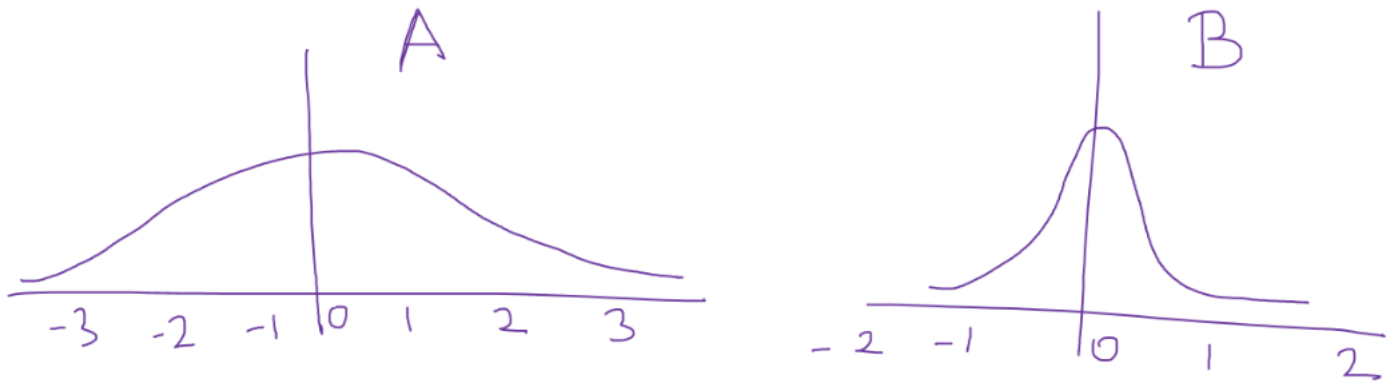
Variance for a sample is calculated as:

$$\nu = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2$$

STANDARD DEVIATION

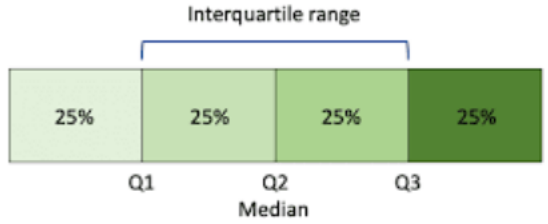
Standard Deviation is the square root of the variance. We square the term $x_i - \bar{x}$ so that we do not care the negative or the positive variation from the mean. We are measuring the average variation from the mean of all the data points. Standard deviation $\sigma = \sqrt{\mu}$, and thus gives a more reasonable estimate to understand the variation.

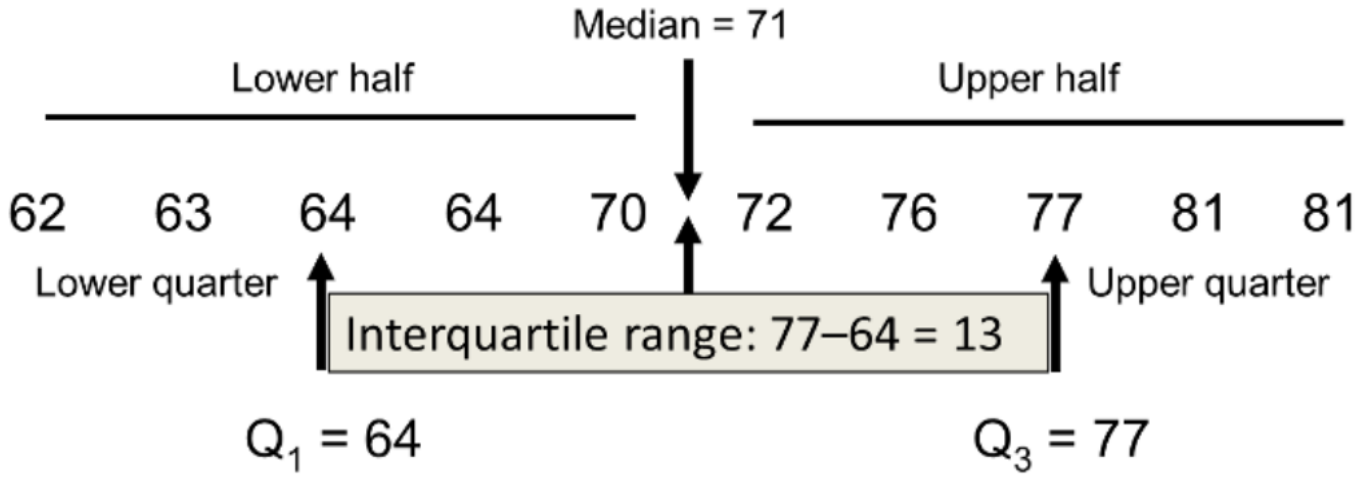
Standard Deviation Formula	
Population	**Sample**
$\sigma = \sqrt{\frac{\sum(X-\mu)^2}{N}}$	$s = \sqrt{\frac{\sum(X-\bar{x})^2}{n-1}}$
X – The Value in the data distribution *μ – The population Mean* *N – Total Number of Observations*	*X – The Value in the data distribution* *$\bar{x}$ – The Sample Mean* *n - Total Number of Observations*



> In the above figure, which one of them have higher variance, A or B? Why?

IQR (Inter Quartile Range) Median usually gives us 50 percentile. i.e. 50% of values are behind the median and rest 50% are ahead of it. We can similarly find 25 percentile and 75 percentile. The number of values that lie from 25 to 75 percentile is the IQR. Quartile comes from Quarter of 4 parts. We have divided the data in 4 parts, <25%, 25-50%, 50-75% and >75%.





VARIANCE

$$\sigma^2 = \frac{\Sigma(xi - \bar{x})^2}{N}$$

The **variance** *of a random variable* X *is*

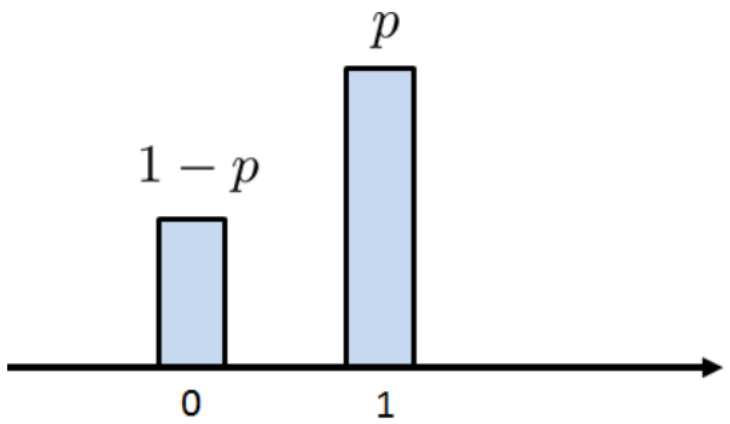
$$\mathrm{Var}[X] = \mathbb{E}[(X - \mu)^2],$$

where $\mu = \mathbb{E}[X]$ *is the expectation of* X.

- variance Var[X] is computed using the ideal histogram PMF.
- What does the variance mean? It is a measure of the deviation of the random variable X relative to its mean

BERNOULI RANDOM VARIABLE

- A Bernoulli random variable is a coin-flip random variable.
- The random variable has two states: either 1 or 0.
- The probability of getting 1 is p, and the probability of getting 0 is 1 − p.



A Bernoulli random variable has two states with probability p and $1-p$.

BERNOULI DISTRIBUTION

- Bernoulli distribution is a discrete probability distribution, meaning it's concerned with discrete random variables. It enables you to calculate the probability of each outcome
- A discrete random variable is one that has a finite or countable number of possible values—
  - the number of heads you get when tossing three coins at once, or the number of students in a class.
- A discrete probability distribution describes the probability that each possible value of a discrete random variable will occur—
  - for example, the probability of getting a six when rolling a die.
- Bernoulli distribution applies to events that have one trial and two possible outcomes.
- These are known as Bernoulli trials.
  - a yes or no question
  - will this coin land on heads when I flip it?
  - Will I roll a six with this die?
  - Will I pick an ace from this deck of cards?
  - Will student Y pass their math test?

BINOMIAL THEOREM

(**Binomial theorem**). *For any real numbers a and b, the binomial series of power n is*

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^{n-k} b^k,$$

where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$.

The **binomial theorem** is valid for any real numbers a and b. The quantity $\binom{n}{k}$ reads as "n choose k". Its definition is

$$\binom{n}{k} \stackrel{\text{def}}{=} \frac{n!}{k!(n-k)!},$$

Find $(1+x)^3$

Using the binomial theorem, we can show that

$$(1+x)^3 = \sum_{k=0}^{n} \binom{3}{k} 1^{3-k} x^k$$
$$= 1 + 3x + 3x^2 + x^3.$$

Let $0 < p < 1$. Find

$$\sum_{k=0}^{n} \binom{n}{k} p^{n-k}(1-p)^k$$

By using the binomial theorem, we have

$$\sum_{k=0}^{n} \binom{n}{k} p^{n-k}(1-p)^k = (p + (1-p))^n = 1.$$

This result will be helpful when evaluating binomial random variables

BINOMIAL RANDOM VARIABLE

- Suppose we flip the coin n times count the number of heads. Since each coin flip is a random variable (Bernoulli), the sum is also a random variable. It turns out that this new random variable is the binomial random variable.

Let X be a **binomial random variable**. *Then, the PMF of X is*

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k = 0, 1, \ldots, n,$$

where $0 < p < 1$ is the binomial parameter, and n is the total number of states. We write

$$X \sim \text{Binomial}(n, p)$$

to say that X is drawn from a binomial distribution with a parameter p of size n.

BINOMIAL RANDOM VARIABLE

- a Binomial RV is a set of Bernoulli experiments or trials.
- The main conditions that need to be fulfilled to define our RV as Binomial:
  - The trials are independent;
  - Each trial can be classified as either success or failure;
  - There is a fixed number of trials;
  - The probability of success on each trial is constant.

- Let's define the RV Z as the number of successes after n trials where P(success) for each trial is p.
- Let's also define Y, a Bernoulli RV with $P(Y=1)=p$ and $P(Y=0)=1-p$.
- Y represents each independent trial that composes Z.

BINOMIAL RANDOM VARIABLE

p^3 $p^2(1-p)$ $p(1-p)^2$ $(1-p)^3$

: The probability of getting k heads out of $n = 3$ coins.

$$p_X(k) = \underbrace{\binom{n}{k}}_{\text{number of combinations}} \underbrace{p^k}_{\text{prob getting } k \text{ H's}} \underbrace{(1-p)^{n-k}}_{\text{prob getting } n-k \text{ T's}}$$

GEOMETRIC SERIES

- A geometric series is the sum of a finite or an infinite sequence of numbers with a constant ratio between successive terms.
- appears naturally in the context of discrete events

Let $0 < r < 1$, a **finite geometric sequence** *of power n is a sequence of numbers*

$$\left\{1, r, r^2, \ldots, r^n\right\}.$$

An **infinite geometric sequence** *is a sequence of numbers*

$$\left\{1, r, r^2, r^3, \ldots\right\}.$$

The sum of a **finite geometric series** *of power n is*

$$\sum_{k=0}^{n} r^k = 1 + r + r^2 + \cdots + r^n = \frac{1 - r^{n+1}}{1 - r}.$$

GEOMETRIC RANDOM VARIABLE

- In some applications, we are interested in trying a binary experiment until we succeed.
- In this case, the random variable can be defined as the outcome of many failures followed by a final success. This is called the geometric random variable.

Let X be a **geometric random variable**. *Then, the PMF of X is*

$$p_X(k) = (1-p)^{k-1}p, \qquad k = 1, 2, \ldots,$$

where $0 < p < 1$ is the geometric parameter. We write

$$X \sim \text{Geometric}(p)$$

$$p_X(k) = \underbrace{(1-p)^{k-1}}_{k-1 \text{ failures}} \underbrace{p}_{\text{final success}}.$$

UNIFORM RANDOM VARIABLE

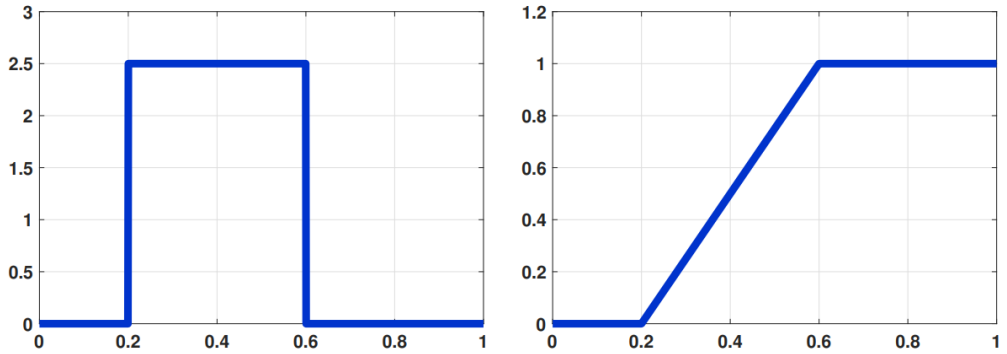
Let X be a continuous **uniform random variable**. *The PDF of X is*

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & a \le x \le b, \\ 0, & \text{otherwise}, \end{cases}$$

where $[a, b]$ is the interval on which X is defined. We write

$$X \sim \text{Uniform}(a, b)$$

to mean that X is drawn from a uniform distribution on an interval $[a, b]$.



(a) PDF (b) CDF

Figure 4.19: The PDF and CDF of $X \sim \text{Uniform}(0.2, 0.6)$.

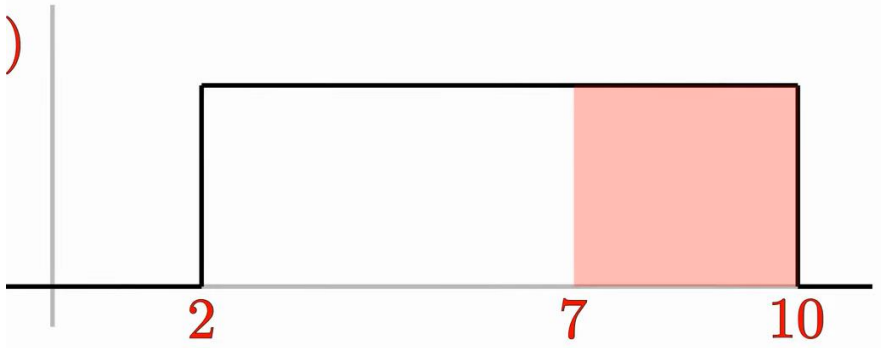
$$\mu = \frac{a+b}{2}$$

$$\sigma = \sqrt{\frac{(b-a)^2}{12}}$$

$$P(c \le X \le d) = \frac{d-c}{b-a}$$

Ex Bus is uniformly late between 2 and 10 minutes. How long can you expect to wait? With what S.D.? If it's >7 mins late, you'll be late for work. What's the prob. of you being late?

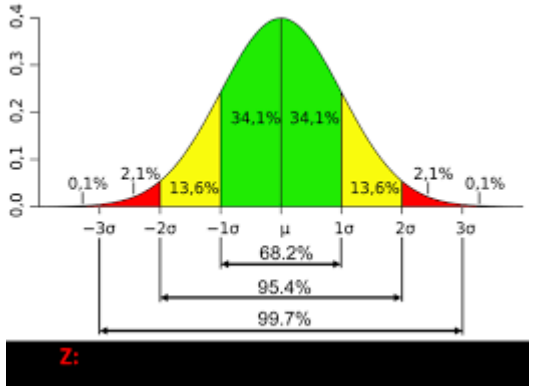
$\mu = 6\ mins$ $\sigma = 2.31\ mins$



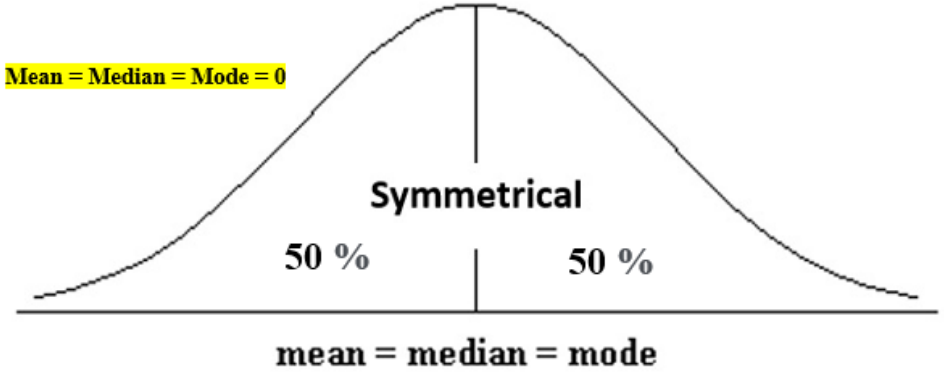
$$P(7 \le X \le 10) = \frac{10-7}{10-2} = 0.375$$

NORMAL DISTRIBUTION

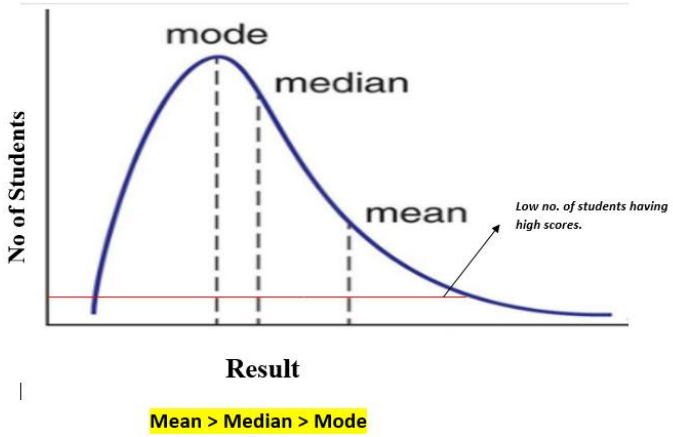
- Normal distribution represents the behavior of most of the situations in the universe
- Any distribution is known as Normal distribution if it has the following characteristics:
  - The mean, median and mode of the distribution coincide.
  - The curve of the distribution is bell-shaped and symmetrical about the line x=μ.
  - The total area under the curve is 1.
  - Exactly half of the values are to the left of the center and the other half to the right.

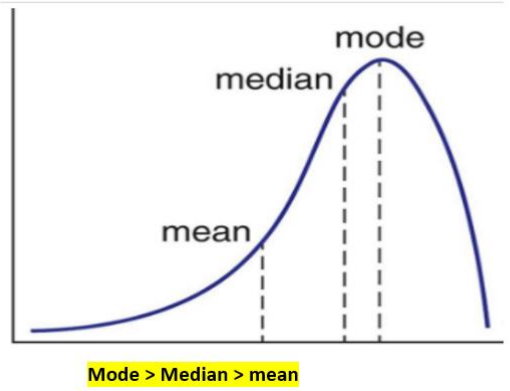


SKEWNESS



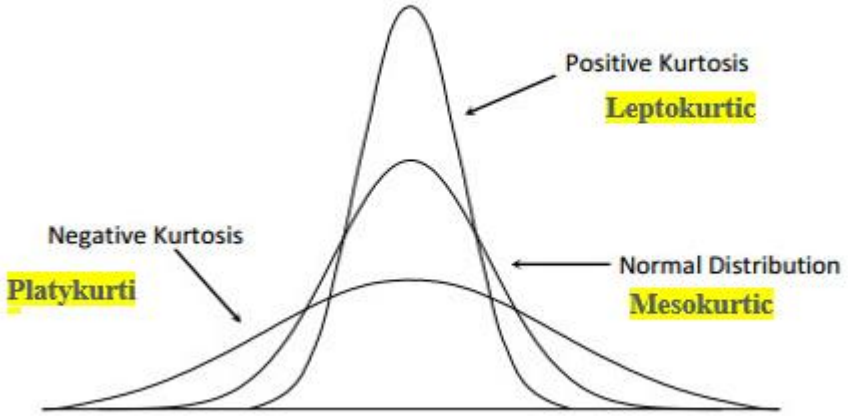
- Skewness essentially measures the symmetry of the distribution, while kurtosis determines the heaviness of the distribution tails.
- When data is symmetrically distributed, the left-hand side, and right-hand side, contain the same number of observations.
- Types of skewness
  - positive and
  - Negative



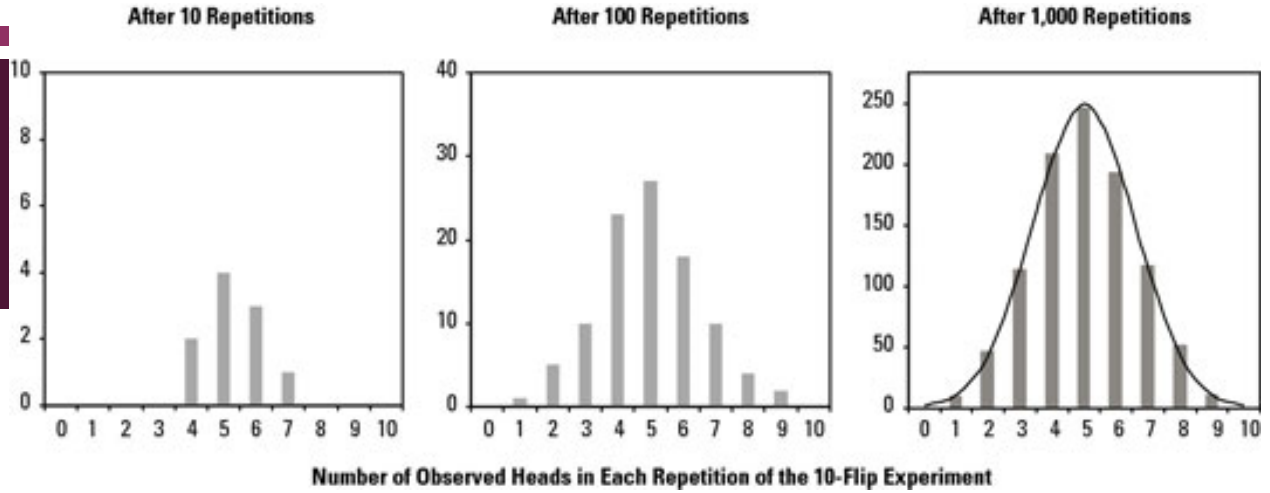


KURTOSIS

- Kurtosis refers to the degree of presence of outliers in the distribution.
- It is a statistical measure, whether the data is heavy-tailed or light-tailed in a normal distribution.
- In finance, kurtosis is used as a measure of financial risk. A large kurtosis is associated with a high level of risk for an investment because it indicates that there are high probabilities of extremely large and extremely small returns. On the other hand, a small kurtosis signals a moderate level of risk because the probabilities of extreme returns are relatively low.



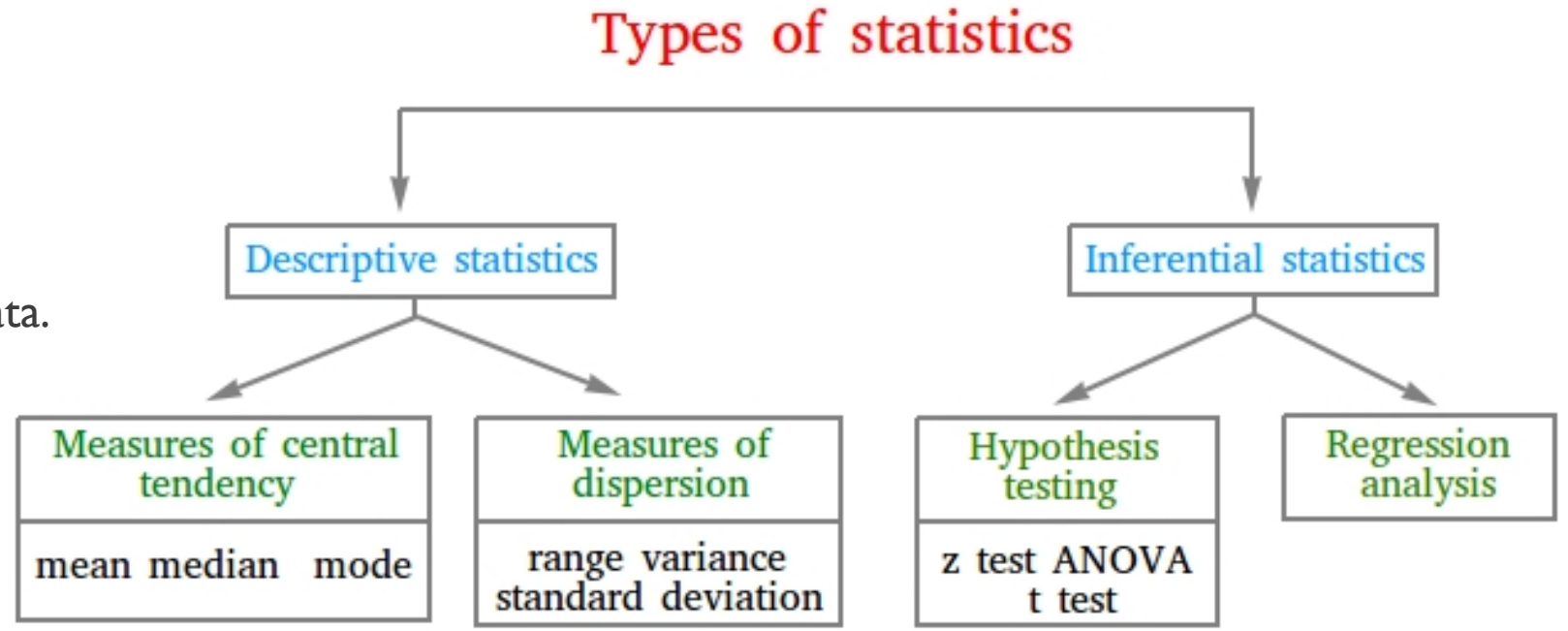
CENTRAL LIMIT THEOREM



- Central Limit Theorem states that whenever a random sample of size n is taken from any distribution with mean and variance, then the sample mean will be approximately normally distributed with mean and variance. The larger the value of the sample size, the better the approximation to the normal.
- Assumptions of Central Limit Theorem
  - The sample should be drawn randomly following the condition of randomization.
  - The samples drawn should be independent of each other. They should not influence the other samples.
  - When the sampling is done without replacement, the sample size shouldn't exceed 10% of the total population.
  - The sample size should be sufficiently large.
  - Refer to jupyter notebook

DESCRIPTIVE STATISTICS

- Describing what is present in the data.

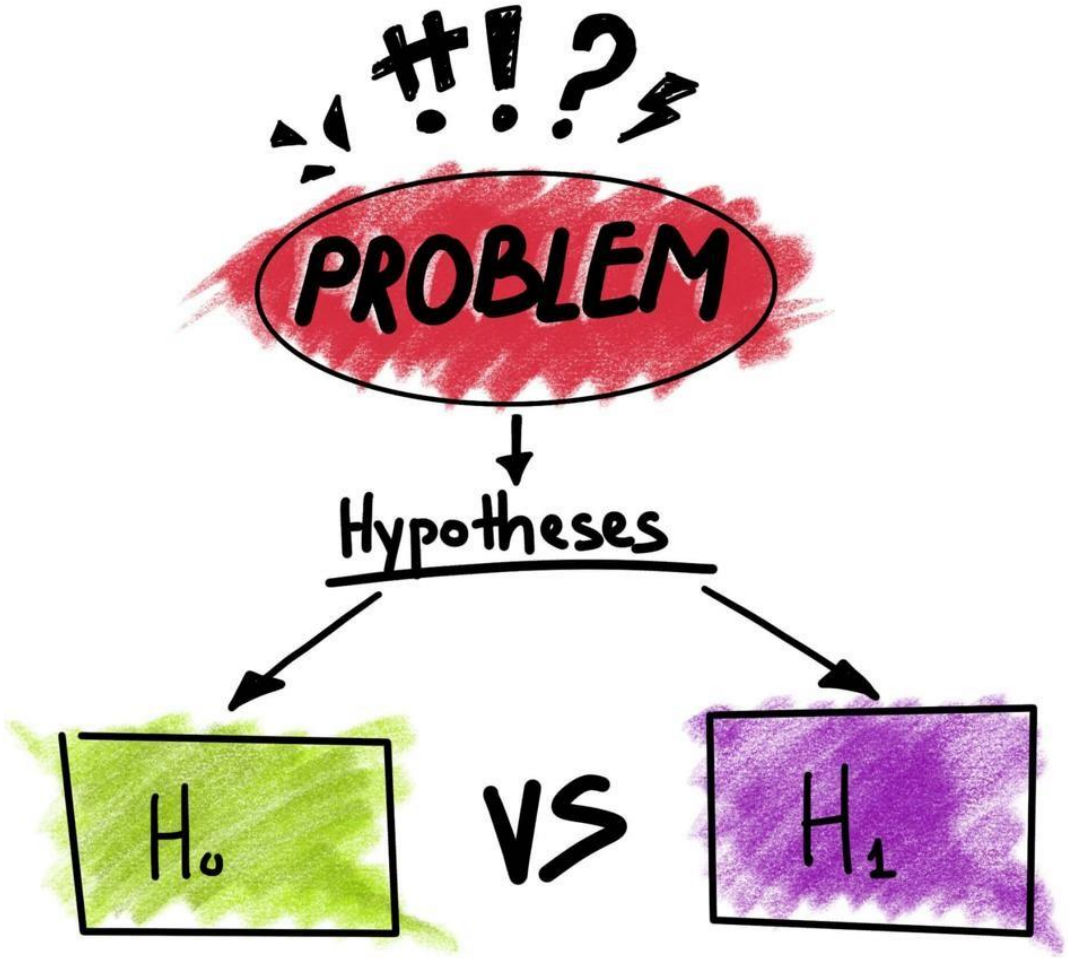


INFERENTIAL STATISTICS

- Makes inferences and predictions about extensive data by considering a sample data from the original data.
- To reach out conclusions that extend beyond the data.
- The process of inferring insights from a sample data is called Inferential statistics
  - Eg: Predicting the amount of rainfall we get in next month by weather forecast.

INFERENTIAL STATISTICS

- Hypothesis Testing comes under Inferential Statistics.
- To do it, we should have the knowledge of
  - Probability
    - Sample Space, Event Space, Random Experiment, Random variable, conditional probability etc.
  - Probability Distribution
  - Sampling Distribution
  - Central Limit Theorem
- Do we know all these?


#!?
PROBLEM
Hypotheses
H₀
VS
H₁

WHAT IS HYPOTHESIS TESTING?

- Hypothesis is a specific testable prediction.
- Hypothesis testing is a part of statistical analysis, where we test the assumptions made regarding a population parameter.
- It is generally used when we were to compare:
  - a single group with an external standard
  - two or more groups with each other
- What is the difference between the terms Parameter and Statistic.
- A Parameter is a number that describes the data from the population whereas, a Statistic is a number that describes the data from a sample.

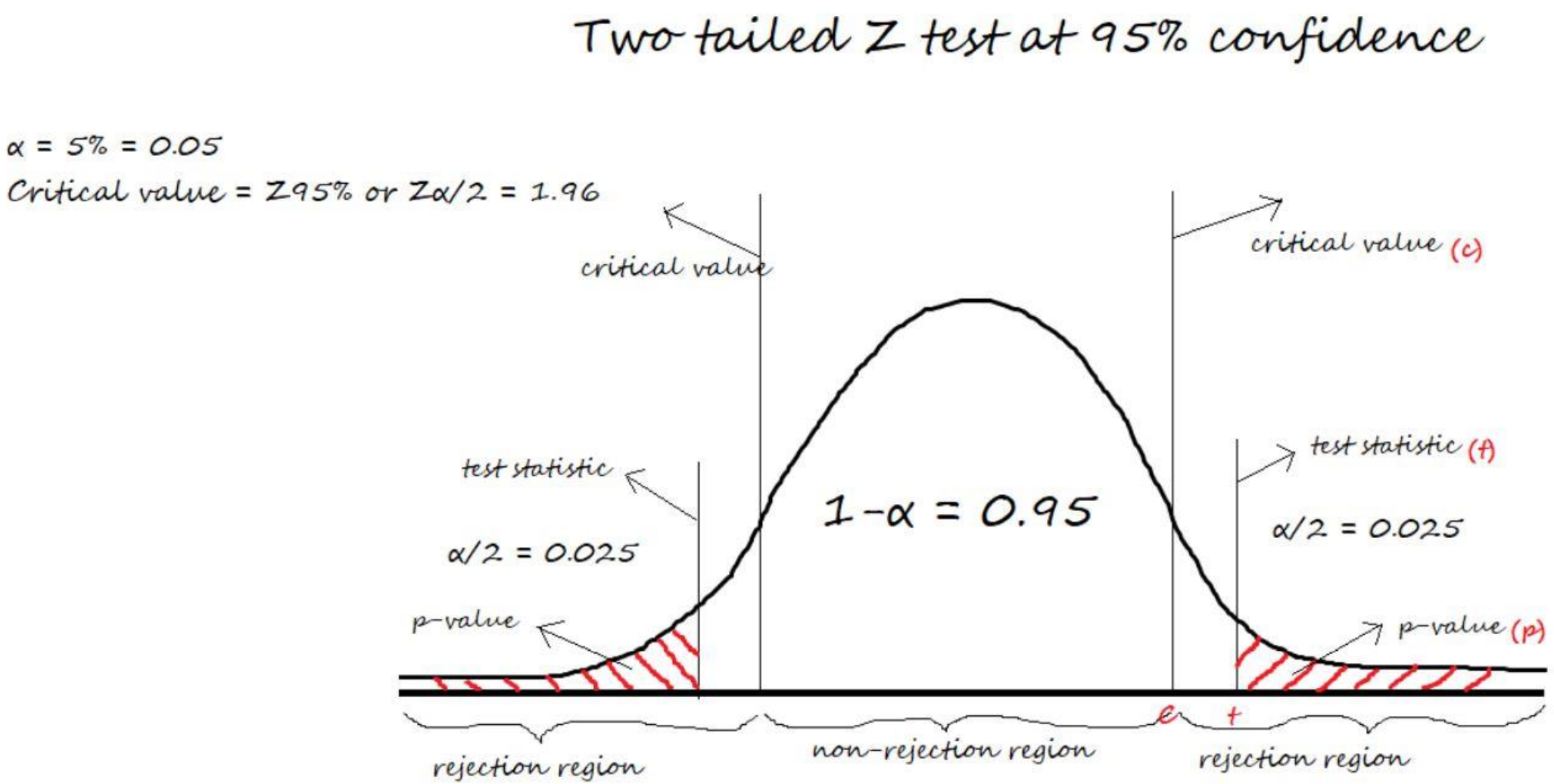
TYPES OF HYPOTHESIS

- Null Hypothesis:
  - Null hypothesis is a statistical theory that suggests there is no statistical significance exists between the populations.
  - It is denoted by H_0
  - A statement of no difference between the means of two populations.
  - Eg: There will be no difference in the performance of students in situations A and B.
- Alternative Hypothesis:
  - An Alternative hypothesis suggests there is a significant difference between the population parameters. It could be greater or smaller. Basically, it is the contrast of the Null Hypothesis.
  - It is denoted by Ha or H1.
  - Eg: Students in situation A will perform better than students in situation B.
- Researchers use inferential statistics to determine the probability that the Null Hypothesis is untrue.

FEW MORE TERMS IN HYPOTHESIS TESTING

- **Level of significance:**
  - Denoted by alpha or α.
  - It is a fixed probability of wrongly rejecting a True Null Hypothesis.
  - For example, if α=5%, that means we are okay to take a 5% risk and conclude there exists a difference when there is no actual difference.
- **Critical Value:**
  - Denoted by C and it is a value in the distribution beyond which leads to the rejection of the Null Hypothesis.
- **Test Statistic:**
  - It is denoted by t and is dependent on the test that we run.
  - It is deciding factor to reject or accept Null Hypothesis.
- **p-value:**
  - It is the proportion of samples (assuming the Null Hypothesis is true) that would be as extreme as the test statistic.
  - It is denoted by the letter p.

- take a look at the below figure for a better understanding of critical value, test-statistic, and p-value

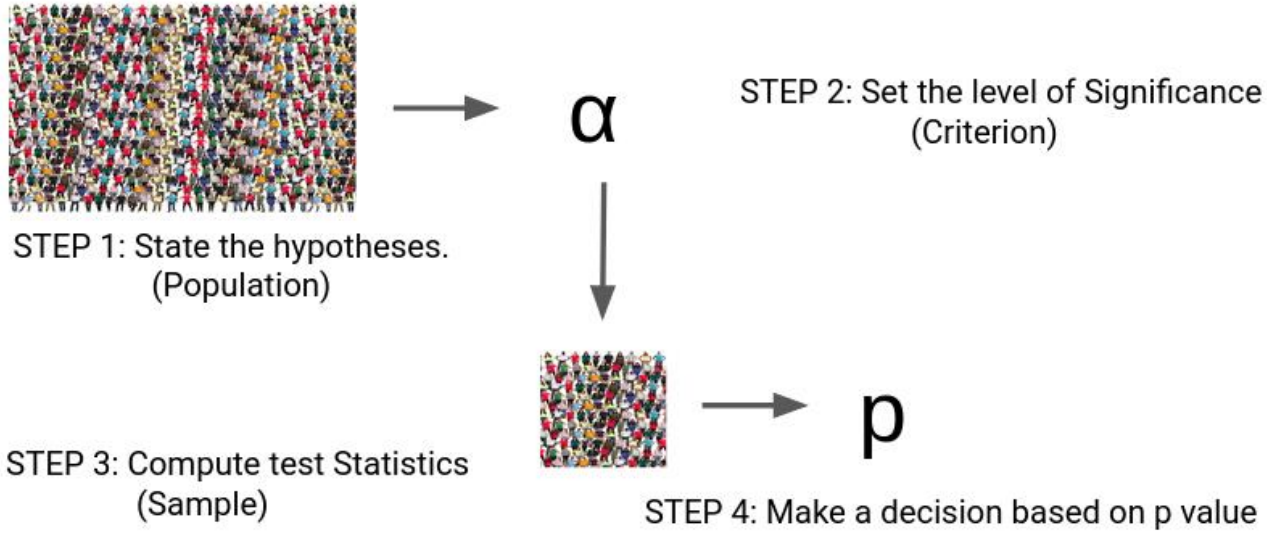


TYPES OF TEST STATISTICS

Hypothesis test	Test Statistic
Z-Test	Z-Score
T-Test	T-Score
F-Test	F-Statisticc
Chi-Square test	Chi-Square Statistic

STEPS OF HYPOTHESIS TESTING

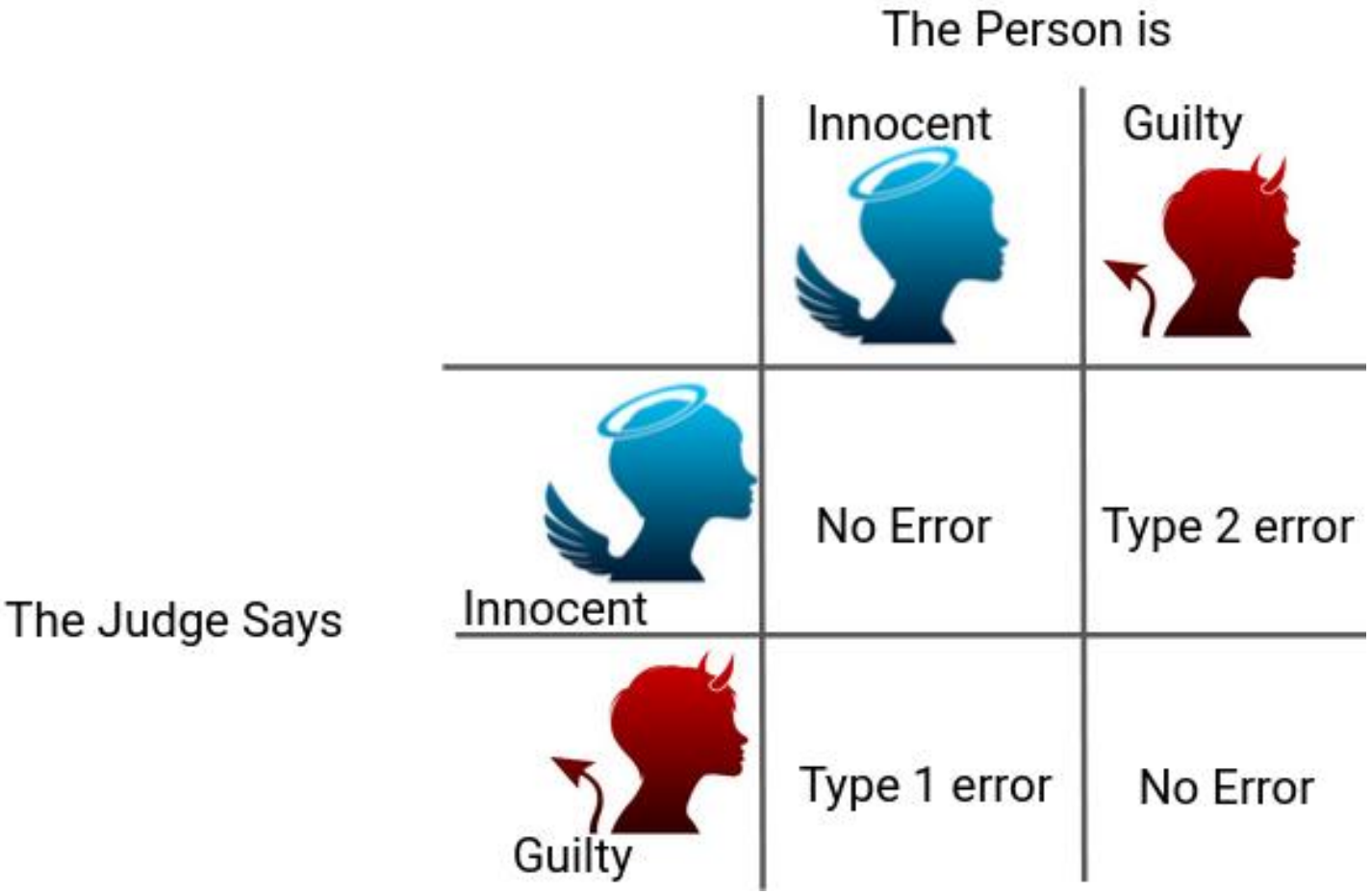
- For a given business problem,
  - Start with specifying Null and Alternative Hypotheses about a population parameter
  - Set the level of significance (α)
  - Collect Sample data and calculate the Test Statistic and P-value by running a Hypothesis test that well suits our data
  - Make Conclusion: Reject or Fail to Reject Null Hypothesis



DECISION RULES

- The two methods of concluding the Hypothesis test are using
  - the Test-statistic value,
  - the p-value.
- In both methods, we start assuming the Null Hypothesis to be true, and then we reject the Null hypothesis if we find enough evidence.
- The decision rule for the Test-statistic method:
  - if test-statistic (t) > critical Value (C), we reject Null Hypothesis.
  - If test-statistic (t) ≤ critical value (C), we fail to reject Null Hypothesis.
- The decision rule for the p-value method:
  - if p-value (p) > level of significance (α), we fail to reject Null Hypothesis
  - if p-value (p) ≤ level of significance (α), we reject Null Hypothesis
- In easy terms, we say **P High, Null Fly** and **P low, Null go**.

- Let’s take an example to understand the concept of Hypothesis Testing. A person is on trial for a criminal offense and the judge needs to provide a verdict on his case.

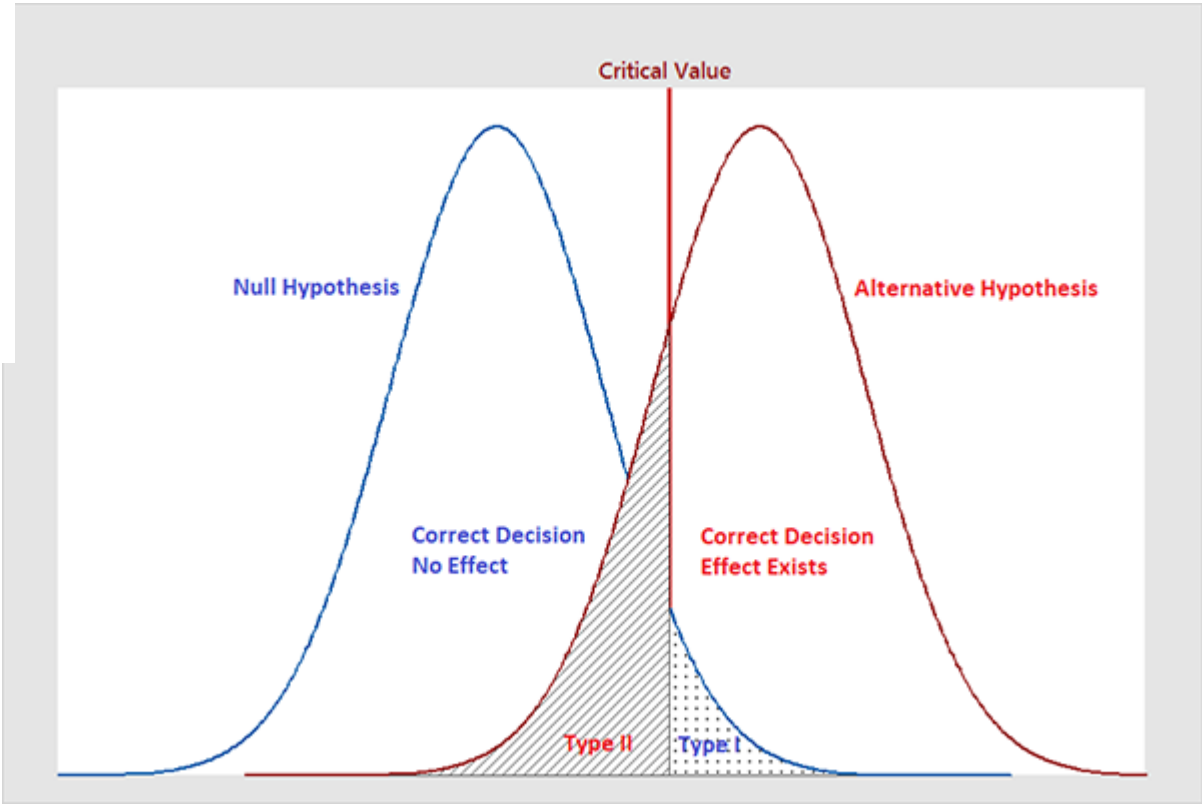


CONFUSION MATRIX IN HYPOTHESIS TESTING

predicted \ Actuals	H0	Ha
Fail to reject H0	correct decision Confidence $(1-\alpha)$	wrong decision Type II error (β)
reject H0	wrong decision Type I error (α)	correct decision Power of the test $(1-\beta)$

$$\text{Accuracy} = \frac{\#\ \text{correct predictions}}{\#\ \text{total cases}}$$

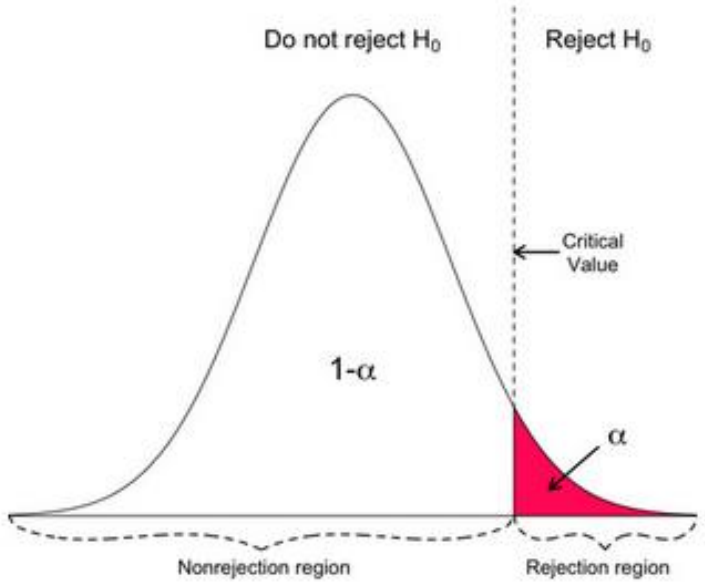
the true state of nature	conclusion from statistical analysis: Accept the Null	conclusion from statistical analysis: Reject the Null
Null hypothesis is **true**	Correct Conclusion	Type I Error reject a true null hypothesis
Null hypothesis is **false**	Type II Error accept a false null hypothesis	Correct Conclusion



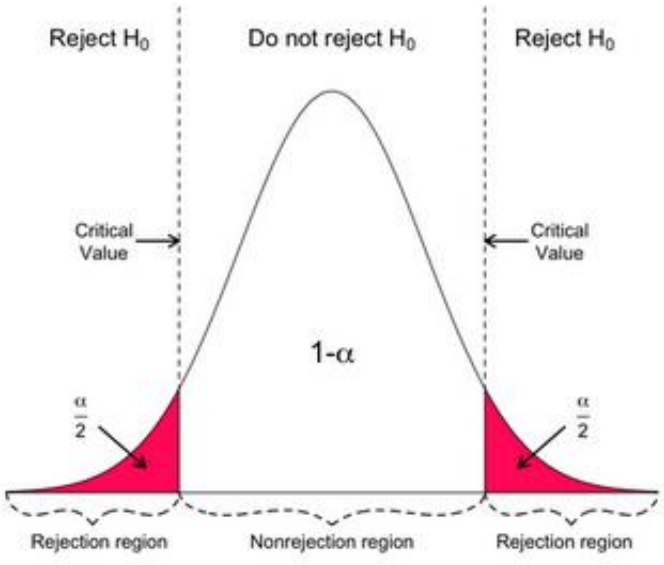
CONFUSION MATRIX

- **Confidence**: The probability of accepting a True Null Hypothesis. It is denoted as (1-α)
- **Power of test:** The probability of rejecting a False Null Hypothesis i.e., the ability of the test to detect a difference. It is denoted as (1-β) and its value lies between 0 and 1.
- **Type I error**: Occurs when we reject a True Null Hypothesis and is denoted as α.
- **Type II error**: Occurs when we accept a False Null Hypothesis and is denoted as β.
- **Accuracy:** Number of correct predictions / Total number of cases
- The factors that affect the power of the test are sample size, population variability, and the confidence (α).
- Confidence and power of test are directly proportional. Increasing the confidence increases the power of the test.

DIRECTIONAL AND NON-DIRECTIONAL HYPOTHESIS



Ex. Females will score significantly higher than Males on IQ score



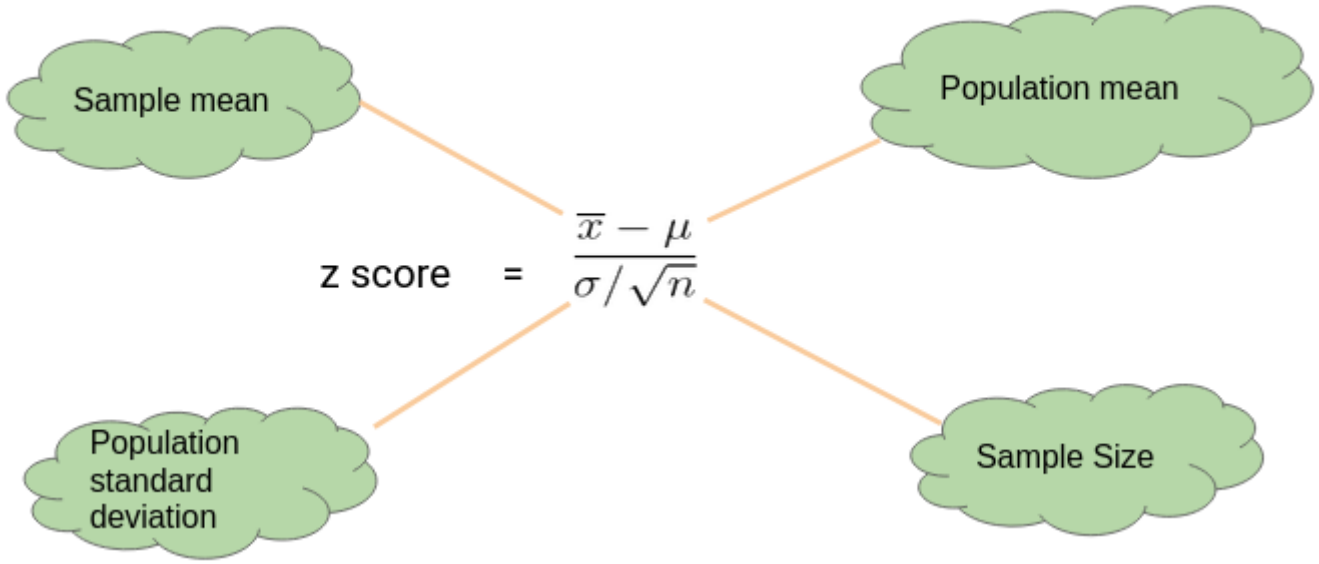
Ex. Females will score significantly different than Males on IQ score

WHAT IS Z-TEST

- z tests are a statistical way of testing a hypothesis when either:
  - We know the population variance, or
  - We do not know the population variance but our sample size is large $n \geq 30$
- If we have a sample size of less than 30 and do not know the population variance, then we must use a t-test.

ONE-SAMPLE Z-TEST

- We perform the One-Sample Z test when we want to compare a sample mean with the population mean.



EXAMPLE OF ONE SAMPLE Z-TEST

- Let's say we need to determine if girls on average score higher than 600 in the exam. We have the information that the standard deviation for girls' scores is 100. So, we collect the data of 20 girls by using random samples and record their marks. Finally, we also set our α value (significance level) to be 0.05.

Score
650
730
510
670
480
800
690
530
590
620
710
670
640
780
650
490
800
600
510
700

In this example:

- Mean Score for Girls is 641
- The size of the sample is 20
- The population mean is 600
- Standard Deviation for Population is 100

$$\text{z score} = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}} = \frac{641 - 600}{100/\sqrt{20}} = 1.8336$$

p value = .033357.

Critical Value = 1.645

Z score > Critical Value

P value < 0.05

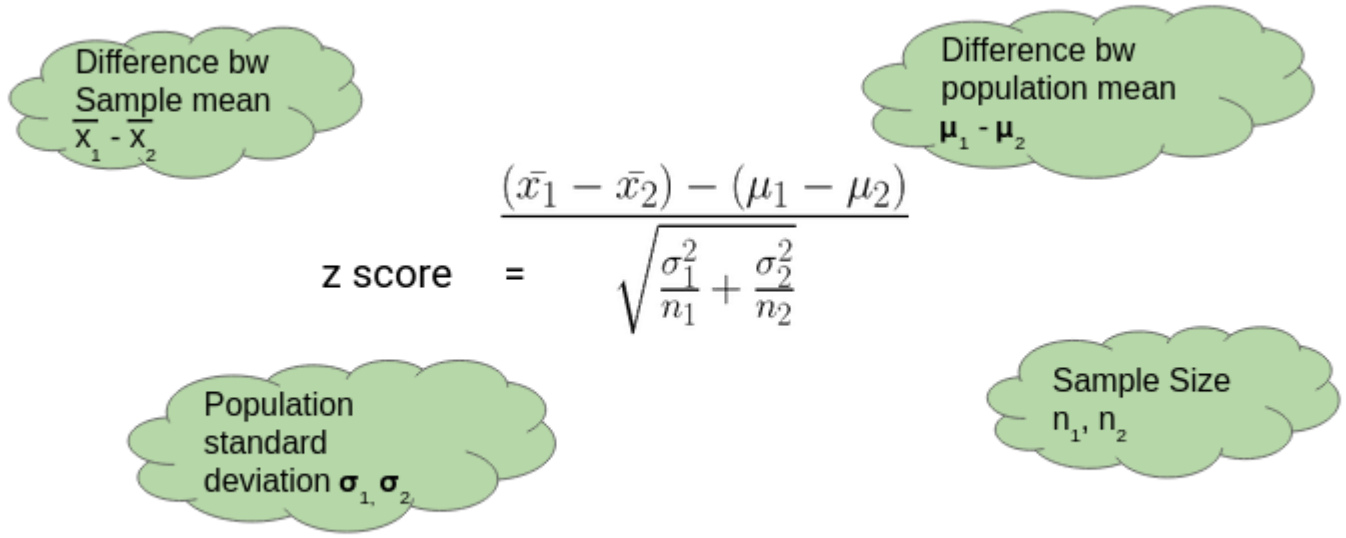
$H_0 : \mu \leq 600$

$H_1 : \mu > 600$

Since the P-value is less than 0.05, we can reject the null hypothesis and conclude based on our result that Girls on average scored higher than 600.

TWO SAMPLE Z TEST

- We perform a Two Sample Z test when we want to compare the mean of two samples.



EXAMPLE OF TWO SAMPLE Z-TEST

- we want to know if Girls on average score 10 marks more than the boys. We have the information that the standard deviation for girls' Score is 100 and for boys' score is 90. Then we collect the data of 20 girls and 20 boys by using random samples and record their marks. Finally, we also set our α value (significance level) to be 0.05.

Score
650
730
510
670
480
800
690
530
590
620
710
670
640
780
650
490
800
600
510
700

Score
630
720
462
631
440
783
673
519
543
579
677
649
632
768
615
463
781
563
488
650

In this example:

- Mean Score for Girls (Sample Mean) is 641
- Mean Score for Boys (Sample Mean) is 613.3
- Standard Deviation for the Population of Girls' is 100
- Standard deviation for the Population of Boys' is 90
- Sample Size is 20 for both Girls and Boys
- Difference between Mean of Population is 10

$$\text{z score} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$= \frac{(641 - 613.3) - (10)}{\sqrt{\frac{100^2}{20} + \frac{90^2}{20}}}$$

$$= 0.588$$

P value = 0.278

Critical Value = 1.645

Z score < Critical Value

P value > 0.05

$H_0 : \mu_1 - \mu_2 \leq 10$

$H_1 : \mu_1 - \mu_2 > 10$

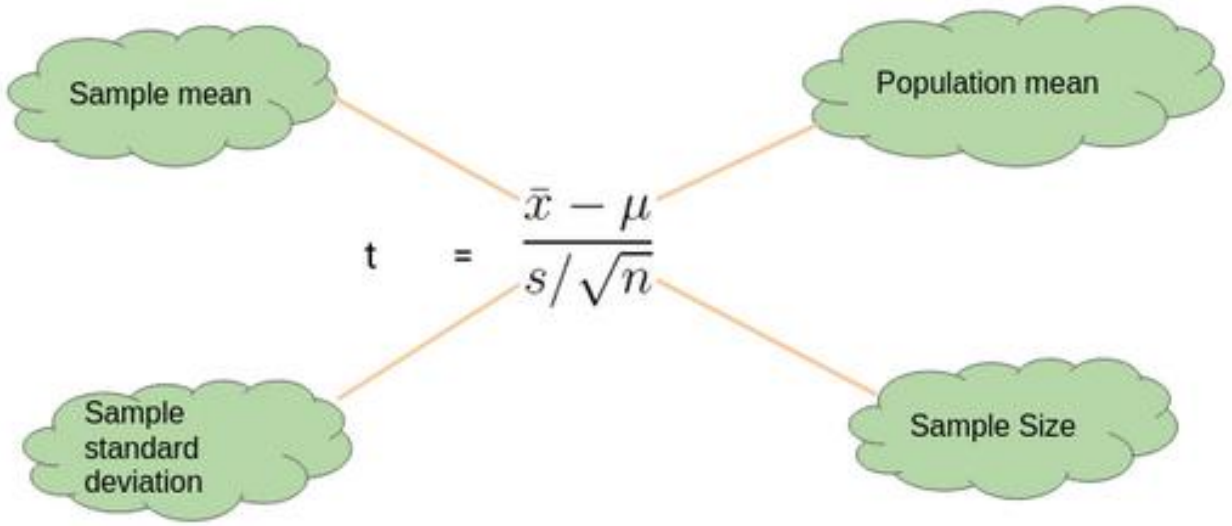
Thus, we can **conclude based on the P-value that we fail to reject the Null Hypothesis**. We don't have enough evidence to conclude that girls on average score of 10 marks more than the boys. Pretty simple, right?

WHAT IS T-TEST?

- t-tests are a statistical way of testing a hypothesis when:
  - We do not know the population variance
  - Our sample size is small, $n < 30$

ONE SAMPLE T-TEST

- We perform a One-Sample t-test when we want to compare a sample mean with the population mean. The difference from the Z Test is that we do not have the information on Population Variance here. We use the sample standard deviation instead of population standard deviation in this case



EXAMPLE TO UNDERSTAND ONE SAMPLE T-TEST

- Let's say we want to determine if on average girls score more than 600 in the exam. We do not have the information related to variance (or standard deviation) for girls' scores. To a perform t-test, we randomly collect the data of 10 girls with their marks and choose our α value (significance level) to be 0.05 for Hypothesis Testing.

Girls_Score
587
602
627
610
619
622
605
608
596
592

In this example:

- Mean Score for Girls is 606.8
- The size of the sample is 10
- The population mean is 600
- Standard Deviation for the sample is 13.14

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

$$= \frac{606.8 - 600}{13.14/\sqrt{10}}$$

$$= 1.64$$

Critical Value = 1.833

t score < Critical Value

P value = 0.0678

P value > 0.05

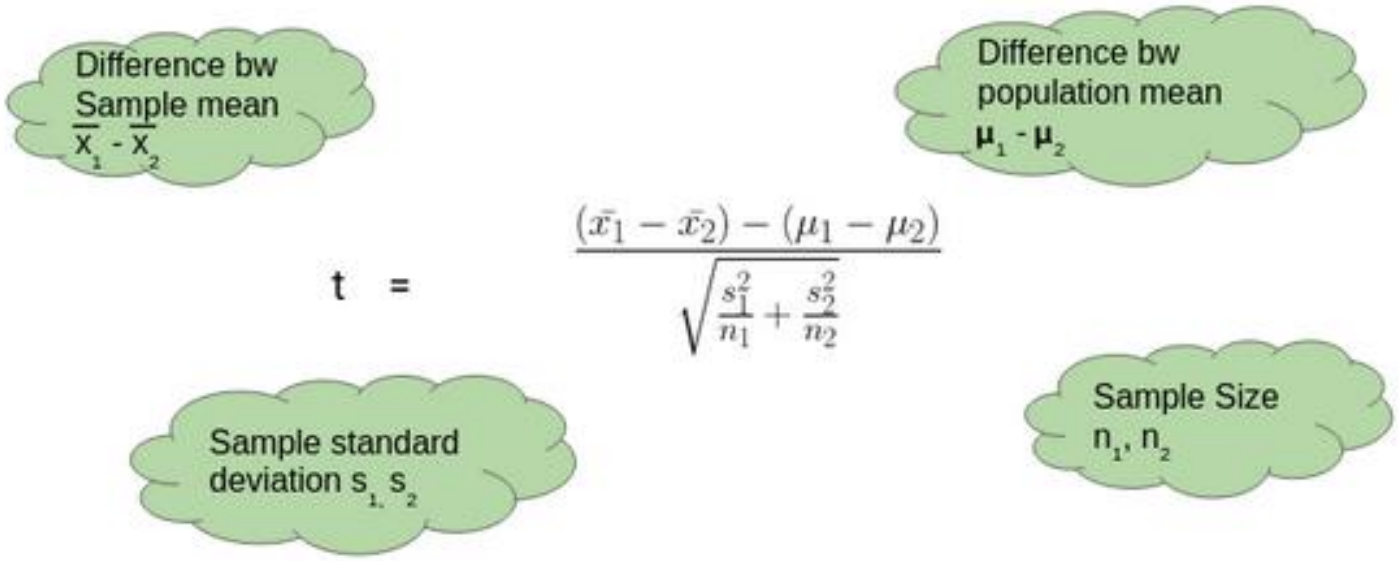
$H_0 : \mu \leq 600$

$H_1 : \mu > 600$

Our **P-value is greater than 0.05 thus we fail to reject the null hypothesis** and don't have enough evidence to support the hypothesis that on average, girls score more than 600 in the exam

TWO SAMPLE T-TEST

- We perform a Two-Sample t-test when we want to compare the mean of two samples.



EXAMPLE TO UNDERSTAND TWO SAMPLE T-TEST

- we want to determine if on average, boys score 15 marks more than girls in the exam. We do not have the information related to variance (or standard deviation) for girls' scores or boys' scores. To perform a t-test. we randomly collect the data of 10 girls and boys with their marks. We choose our α value (significance level) to be 0.05 as the criteria for Hypothesis Testing.

Girls_Score
587
602
627
610
619
622
605
608
596
592

Boys_Score
626
643
647
634
630
649
625
623
617
607

In this example:

- Mean Score for Boys is 630.1
- Mean Score for Girls is 606.8
- Difference between Population Mean 15
- Standard Deviation for Boys' score is 13.42
- Standard Deviation for Girls' score is 13.14

$$t = \frac{(\bar{x_1} - \bar{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$\frac{(630.1 - 606.8) - (15)}{\sqrt{\frac{(13.42)^2}{10} + \frac{(13.14)^2}{10}}}$$

Critical Value = 1.833

t = 2.23

P value = 0.019

Critical Value > t score

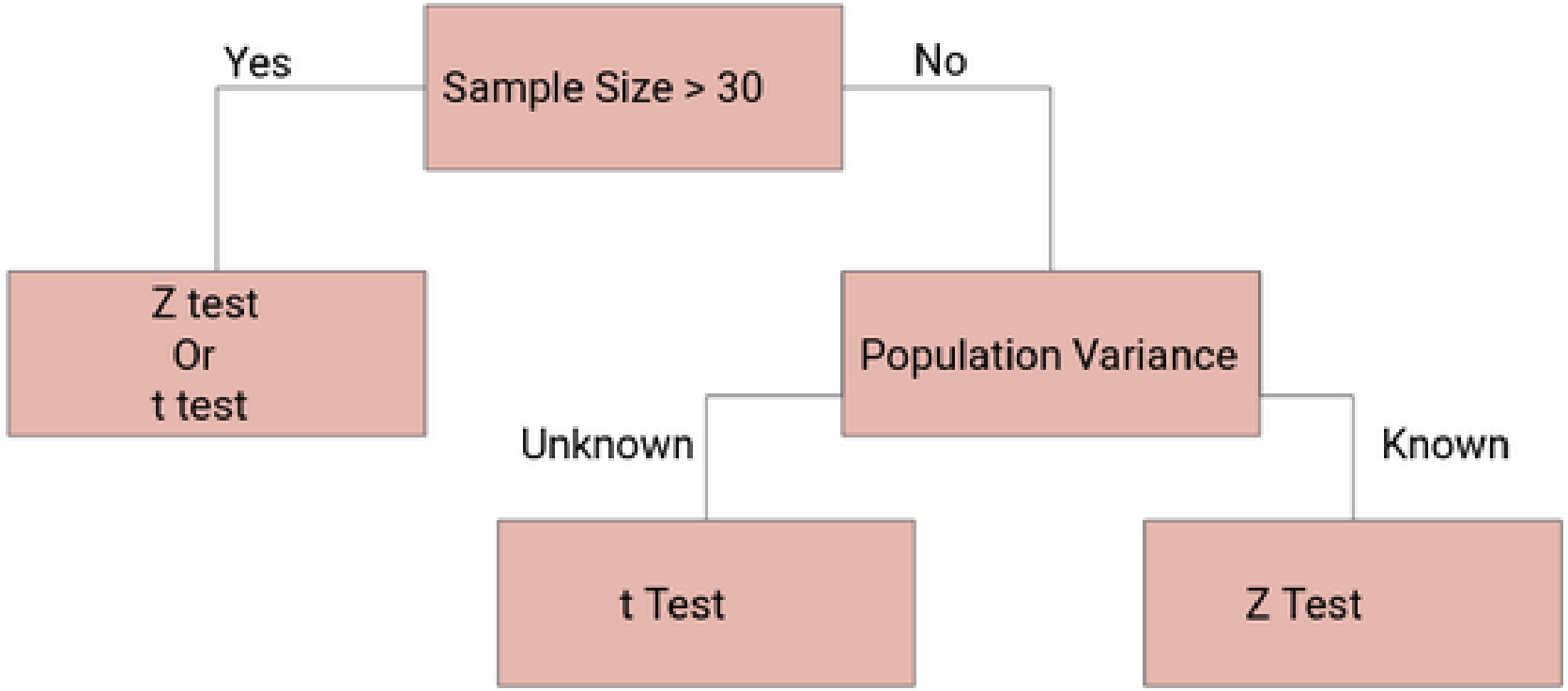
P value < 0.05

$H_0 : \mu_1 - \mu_2 \leq 10$

$H_1 : \mu_1 - \mu_2 > 10$

Thus, **P-value is less than 0.05 so we can reject the null hypothesis** and conclude that on average boys score 15 marks more than girls in the exam.

DECIDING BETWEEN Z-TEST AND T-TEST



TWO SAMPLE T-TEST

Data:

Compare the effectiveness of ammonium chloride and urea, on the grain yield of paddy, an experiment was conducted. The results are given below:

Ammonium chloride (X_1)	13.4	10.9	11.2	11.8	14	15.3	14.2	12.6	17	16.2	16.5	15.7
Urea (X_2)	12	11.7	10.7	11.2	14.8	14.4	13.9	13.7	16.9	16	15.6	16

Hypothesis

H_0: The effect of ammonium chloride and urea on grain yield of paddy are equal i.e., $\mu_1 = \mu_2$

H_1: The effect of ammonium chloride and urea on grain yield of paddy is not equal i.e., $\mu_1 \neq \mu_2$

Refer Colab Notebook

ANOVA TEST

- To compare several population means.
- The basic idea behind a one-way ANOVA is to take independent random samples from each group, then compute the sample means for each group. After that compare the variation of sample means among the groups to the variation within the groups. Finally, make a decision based on a test statistic, whether the means of the groups are all equal or not.
- **Sum of Squares (SS)**
- Inside the One-Way ANOVA Table:
  The total amount of variability comes from two possible sources, namely:
  1. Difference **among** the groups, called **treatment** (TR)
  2. Difference **within** the groups, called **error** (E)
- The sum of the squares due to treatment (**SSTR**) and the sum of squares due to error (**SSE**) are listed in the one-way ANOVA table. The sum of SSTR and SSE is equal to the total sum of squares (**SSTO**).
- **Mean Squares (MS)**
- A mean square is the sum of squares divided by its d.f. These mean squares are all variances and will be used in the hypothesis test of the equality of all the group population means.

ASSUMPTIONS FOR THE ONE-WAY ANOVA HYPOTHESIS TEST

- Sample data are randomly selected from populations and randomly assigned to each of the treatment groups. Each observation is thus independent of any other observation — randomness and independence.
- Normality. Values in each sampled groups are assumed to be drawn from normally distributed populations.
- Homogeneity of variance. All the c group variances are equal, that is $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \ldots = \sigma c^2$.

The simple outline of the one-way ANOVA test:

F test for differences in more than two means

H_0: $\mu_1 = \mu_2 = \mu_3 = \ldots = \mu c$

H_1: Not all μ_i's are equal, where i = 1, 2, 3, ..., c.

Level of significance = α

The test statistic $= F = \dfrac{\text{MSTR}}{\text{MSE}} \sim F_{c-1,n-c}$

Decision Rule: Reject H_0 when $F > F_{\alpha;\, c-1,n-c}$ OR when test $p - value < \alpha$

Finally, the one-way ANOVA table is as shown below:

Source of Variation	d.f.	SS	MS	F	P-value	F crit
Between Groups	c - 1	SSTR	MSTR	$F = \dfrac{\text{MSTR}}{\text{MSE}}$	p	$F_{\alpha;\, c-1,n-c}$
Within Groups	n - c	SSE	MSE			
Total	n - 1	SSTO				

CHI-SQUARE TEST

- The Research Hypothesis (H1) proposes that the two variables are related in the population.
- The Null Hypothesis (H0) states that no association exists between the two variables in the population, and therefore the variables are statistically independent.
- Expected Frequencies : is the cell frequencies that would be expected in a table if the two tables were statistically independent.
- Observed frequencies : is the cell frequencies actually observed in a table.

$$\text{Expected Value} = \frac{(Row\ Total) * (Column\ Total)}{Total\ Number\ of\ Observatiosn}$$

- To obtain the expected frequencies for any cell in which the two variables are assumed independent, multiply the row and column totals for that cell and divide the product by the total number of cases in the table.

CHI-SQUARE TEST

- Chi-Square test is an inferential statistics technique designed to test for significant relationships between two variables.
- Chi-square required no assumptions about the shape of the population distribution from which a sample is drawn.
- Like all inferential statistics, it also assumes random sampling.
- Limitations:
  - The chi-square test does not give us much information about the strength of the relationship.
  - This test is sensitive to sample size. The size of the calculated chi-square is directly proportional to the size of the sample, independent of the strength of the relationship between the variables.

CHI-SQUARE TEST

$$\chi^2 = \sum \frac{(f_e - f_o)^2}{f_e}$$

f_e = expected frequencies

f_o = observed frequencies

Chi-square follows five steps:

- Making assumptions (random sampling)
- Stating the research and null hypotheses
- Selecting the sampling distribution and specifying the test statistic
- Computing the test statistics
- Deciding and interpreting the result

EXAMPLE

The below table gives the relationship between handedness and gender in the U.S.A. This data is also known as HANES data

	Men	Women	Total
Right-Handed	934	1,070	2,004
Left-Handed	113	92	205
Ambidextrous	20	8	28
Total	1,067	1,170	2,237

$$x^2 = sum\ of\ \frac{(observed\ frequency - expected\ frequency)^2}{expected\ frequency}$$

- Null Hypothesis : Gender and handedness are independent

$$\text{Expected Value} = \frac{(Row\ Total) * (Column\ Total)}{Total\ Number\ of\ Observations}$$

$$= \frac{1067 * 2004}{2237} = 956$$

$$\frac{(934-956)^2}{956} + \frac{(1{,}070-1{,}048)^2}{1{,}048} + \frac{(113-98)^2}{98}$$

$$+ \frac{(92-107)^2}{107} + \frac{(20-13)^2}{13} + \frac{(8-15)^2}{15}$$

$$= 12$$

Critical values of the Chi-square distribution with *d* degrees of freedom

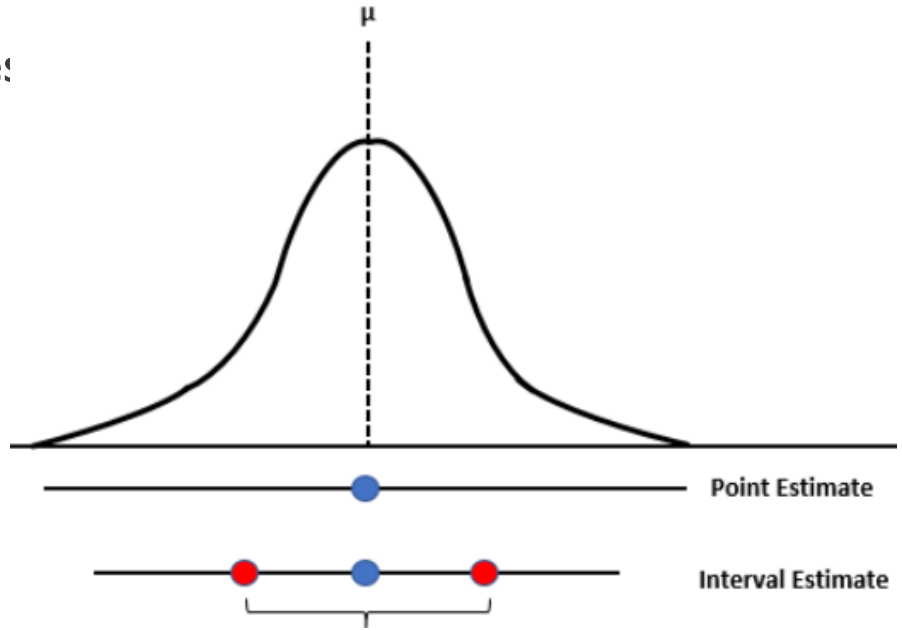
	Probability of exceeding the critical value						
d	0.05	0.01	0.001	*d*	0.05	0.01	0.001
1	3.841	6.635	10.828	11	19.675	24.725	31.264
2	5.991	9.210	13.816	12	21.026	26.217	32.910
3	7.815	11.345	16.266	13	22.362	27.688	34.528
4	9.488	13.277	18.467	14	23.685	29.141	36.123
5	11.070	15.086	20.515	15	24.996	30.578	37.697
6	12.592	16.812	22.458	16	26.296	32.000	39.252
7	14.067	18.475	24.322	17	27.587	33.409	40.790
8	15.507	20.090	26.125	18	28.869	34.805	42.312
9	16.919	21.666	27.877	19	30.144	36.191	43.820
10	18.307	23.209	29.588	20	31.410	37.566	45.315

Chi-squared = 12
Degrees of freedom = 2
Alpha value = 0.05

The Null hypothesis is rejected: Based on the HANES data, handedness and gender are very likely not independent

ESTIMATORS

- Estimation is a process in which we obtain the values of unknown population parameters with the help of sample data.
- Using descriptive and inferential statistics, you can make two types of estimates about the population: **point** estimates and **interval** estimates.
- A point estimate is a single value estimate of a parameter. For instance, a sample mean is a point estimate of a population mean.
- An interval estimate gives you a range of values where the parameter is expected to lie. A confidence interval is the most common type of interval estimate.
- Both types of estimates are important for gathering a clear idea of where a parameter is likely to lie.



PROPERTIES OF ESTIMATORS

- Sample measures are used to estimate the population measures; these statistics are the estimators. Following are the properties of good estimators.
  - An estimator should be consistent. For instance, if it is consistent, the estimator value approaches the parameter value estimated as the sample size increases.
  - Estimators should be unbiased. In other words, the expected value obtained from the sample is equal to the parameter being estimated. Otherwise, the estimator is biased.
  - The estimator should be efficient. In other words, it should have minimal variance to the actual variance of the estimator.

VARIABLES

- x: The individual value
- $\bar{X}$: a point estimate for the population mean
- σ: the actual population standard deviation / symbol for the measurement of dispersion in a population
- n: The statistic for number of data in a sample
- N is for populations
- $\bar{\bar{X}}$:(double bar): The grand average of the subgroup averages. AKA X-bar bar or X-double bar
- s (or sd): The sample standard deviation is a point estimate for the population standard deviation / the dispersion statistic for samples
- μ: the central tendency statistic for populations

TYPES OF ESTIMATES

- Point Estimates
  - Point estimate is a single value derived from a sample and used to estimate the population value.
  - For example, 62 is the average ($\bar{x}$) marks achieved by a sample of 15 students randomly collected from a class of 150 students is considered to be the mean marks of the entire class.
  - The basic drawback of point estimate is that no information is available regarding the reliability.
- Interval Estimates
  - A confidence interval estimate is a range of values constructed from sample data so that the population parameter is likely to occur within the range at a specified probability. The specified probability is the level of confidence.
  - Broader and probably more accurate than a point estimate
  - Used with inferential statistics to develop a confidence interval – where we believe with a certain degree of confidence that the population parameter lies.
  - Any parameter estimate that is based on a sample statistic has some amount of sampling error.

CONFIDENCE INTERVAL

- Confidence interval is to express the precision and ambiguity related to a particular sampling method. Additionally, the confidence interval equation consists of 3 parts.
- A confidence interval is a range of values that probably contain the population mean.
- Confidence level is a percentage of certainty that in any given sample, that confidence interval will contain the population mean.
- Point estimate is a statistic (value from a sample) is to estimate a parameter (value from the population).
- Margin of error is the maximum expected difference between the actual population parameter and a sample estimate of the parameter. In other words, it is the range of values above and below sample statistics.