P P Savani University
School of Engineering

Institute of Computer Science and Application

SSCA3021: Data Science

Module 3: Data Analysis & Machine Learning Algorithms

PPSU
P P SAVANI UNIVERSITY

By Misha Patel

Terminology and Concepts of Data Analysis

Introduction to Data Analysis:

- **What is data analysis?** The process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information, informing conclusions, and supporting decision-making.
- **Why is it important?** Highlighting its role in various fields (business, science, healthcare, etc.) for gaining insights, making data-driven decisions, identifying trends, and solving problems.
- **Types of data analysis:** descriptive, diagnostic, predictive, and prescriptive analysis.
- **The data analysis process** is often depicted as a cycle involving data collection, cleaning, exploration, modeling, and interpretation.

Terminology and Concepts of Data Analysis:

This part introduces fundamental vocabulary and ideas, such as:

- **Data:** Raw facts, figures, or symbols collected for analysis.
- **Information:** Data processed, organized, and structured in a given context to make it useful.
- **Knowledge:** The understanding gained from information, leading to insights and actionable intelligence.
- **Dataset:** A collection of related data.
- **Variable:** A characteristic or attribute that can be measured or observed.
- **Observation/Record:** A set of values for all variables for a single entity.
- **Population:** The entire group of individuals or instances about which we want to draw conclusions.
- **Sample:** A subset of the population selected for analysis.

WHAT IS PROBABILITY?

Probability is the measure of how likely an event will occur.

- Probability is the ratio of desired outcomes to total outcomes:

(desired outcomes) / (total outcomes)

- Probabilities of all outcomes always sums to 1

Example:

- On rolling a dice, you get 6 possible outcomes
- Each possibility only has one outcome, so each has a probability of 1/6
- For example, the probability of getting a number '2' on the dice is 1/6

TERMINOLOGIES IN PROBABILITY

Random Experiment

An experiment or a process for which the outcome cannot be predicted with certainty

Sample Space

The entire possible set of outcomes of a random experiment is the sample space (S) of that experiment

Event

One or more outcomes of an experiment. It is a subset of sample space(S)

Terminology and Concepts of Probability

Basic Terminology:

- **Experiment:** An action with an uncertain outcome, like flipping a coin.
- **Outcome:** A single possible result of an experiment (e.g., heads or tails).
- **Sample Space (S):** The collection of all possible outcomes of an experiment.
- **Event (E):** A subset of the sample space, containing specific outcomes you're interested in (e.g., getting heads).
- **Random Event:** An event with multiple possible outcomes where chance determines the result (e.g., rolling a die).
- **Fairness:** When all outcomes in a sample space have an equal chance of happening (e.g., a balanced coin).
- **Mutually Exclusive Events:** Events that cannot occur at the same time (e.g., rolling a 1 and a 2 on a single die).
- **Independent Events:** Events where the outcome of one doesn't affect the probability of the other (e.g., flipping a heads on one coin doesn't affect the outcome of the other).

Introduction to Statistics

Statistics is the foundation upon which data science is built. It equips you with the tools and techniques to extract meaning from data, a crucial skill for data scientists.

Continue...

Here's a breakdown of how statistics plays a vital role in data science:

Understanding the Data:

- **Descriptive Statistics:** This helps you summarize and describe key characteristics of your data. You can calculate measures like mean, median, and standard deviation to understand the "central tendency" and spread of your data points.
- **Data Visualization:** Statistical techniques help create charts and graphs that effectively communicate patterns and trends within the data.

These visualizations make it easier to identify potential relationships and outliers.

Continue...

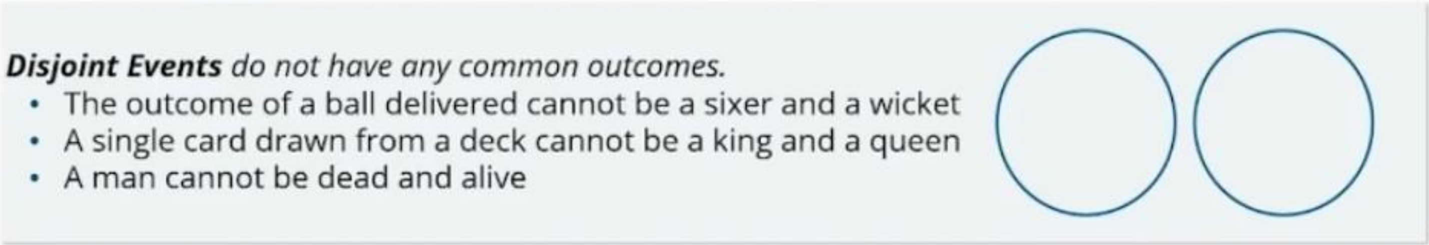
Drawing Inferences:

- **Inferential Statistics:** Allows you to make conclusions about a larger population (population) based on a smaller sample of the data. This is vital, as data science often relies on analyzing samples of the bigger picture.
- **Hypothesis Testing:** Formulate and test hypotheses about the data to see if there's evidence to support them. This helps you move beyond simply describing data to understanding relationships between variables.

Making Predictions:

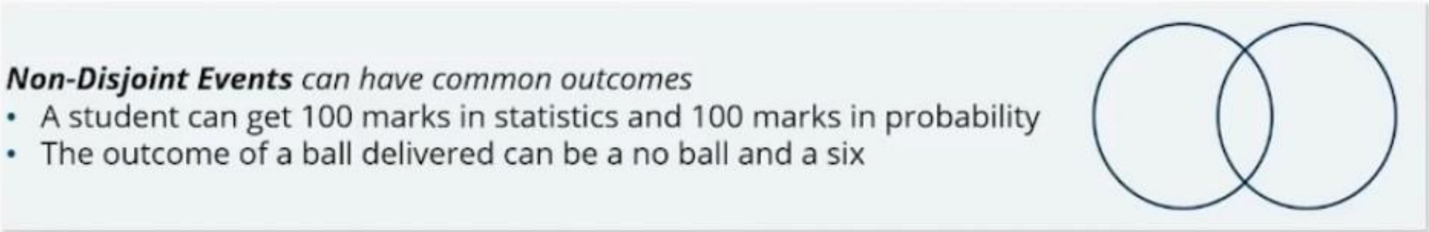
- **Regression Analysis:** Develop models that can predict future outcomes based on historical data. This is a cornerstone of many data science applications.
- **Probability Theory:** Understands the likelihood of events happening, which is crucial for building reliable predictive models.

TYPES OF EVENTS



Disjoint Events *do not have any common outcomes.*

- The outcome of a ball delivered cannot be a sixer and a wicket
- A single card drawn from a deck cannot be a king and a queen
- A man cannot be dead and alive



Non-Disjoint Events *can have common outcomes*

- A student can get 100 marks in statistics and 100 marks in probability
- The outcome of a ball delivered can be a no ball and a six

PROBABILITY DISTRIBUTION

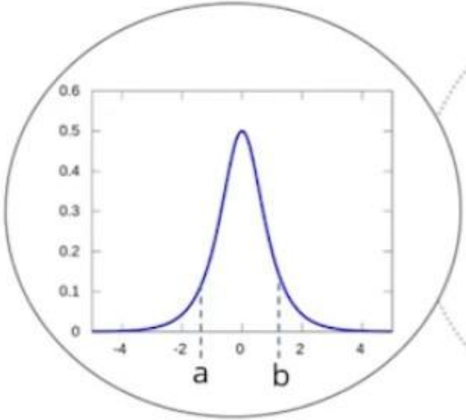
Probability Density Function 01

Normal Distribution 02

Central Limit Theorem 03

PROBABILITY DENSITY FUNCTION

The equation describing a continuous probability distribution is called a Probability Density Function



Property 01

Graph of a PDF will be continuous over a range

Property 02

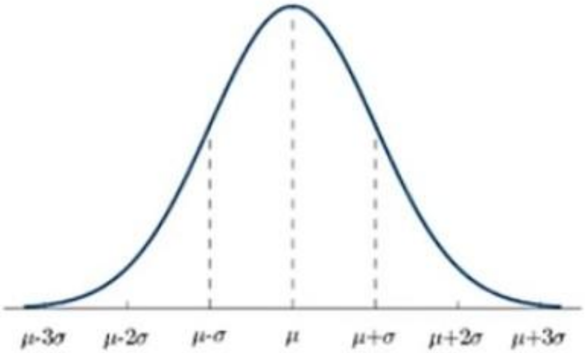
Area bounded by the curve of density function and the x-axis is equal to 1

Property 03

Probability that a random variable assumes a value between a & b is equal to the area under the PDF bounded by a & b

NORMAL DISTRIBUTION

The Normal Distribution is a probability distribution that associates the normal random variable X with a cumulative probability



$$Y = [\ 1/\sigma * \mathrm{sqrt}(2\pi)\ ] * e^{-(x - \mu)2/2\sigma2}$$

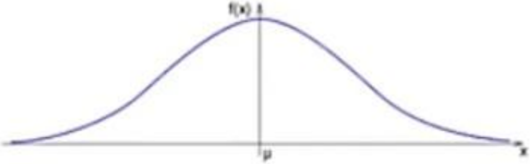
Where,
- *X* is a normal random variable
- μ is the mean and
- σ is the standard deviation

Note: Normal Random variable is variable with mean at 0 and variance equal to 1

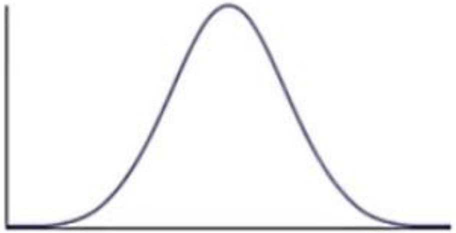
STANDARD DEVIATION & CURVE

The graph of the Normal Distribution depends on two factors: the *Mean* and the *Standard Deviation*

- **Mean:** *Determines the location of center of the graph*
- **Standard Deviation:** *Determines the height of the graph*



If the standard deviation is large, the curve is short and wide.



If the standard deviation is small, the curve is tall and narrow.

Benefits of Statistics in Data Science:

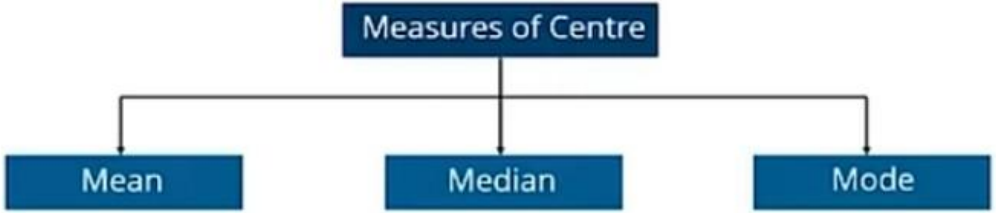
- **Informed Decisions:** Statistics helps you make data-driven decisions that are not based on intuition alone.
- **Uncover Hidden Patterns:** Identify patterns and relationships in complex datasets that might not be readily apparent.
- **Reliable Models:** Statistical techniques ensure the models you build from data are robust and generalizable.
- **Communicate Insights:** Statistics helps you effectively communicate your findings to others using clear and concise language.

DESCRIPTIVE STATISTICS

Descriptive statistics is a method used to describe and understand the features of a specific data set by giving short summaries about the sample and measures of the data.

Descriptive statistics are broken down into two categories:

- **Measures of Central tendency**
- Measures of Variability (spread)



Measures of Centre

Mean

Median

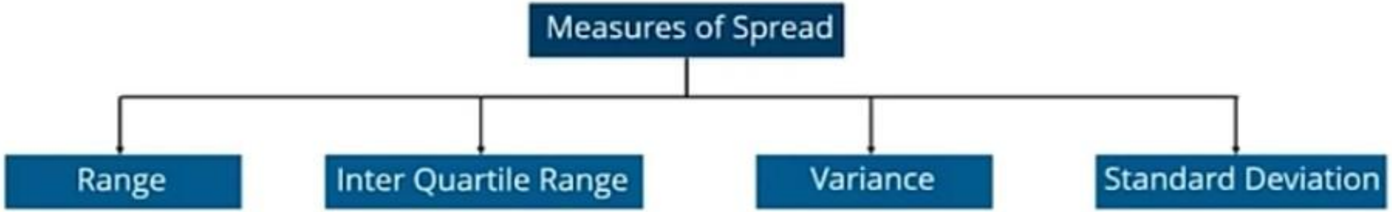
Mode

DESCRIPTIVE STATISTICS

Descriptive statistics is a method used to describe and understand the features of a specific data set by giving short summaries about the sample and measures of the data.

Descriptive statistics are broken down into two categories:

- Measures of Central tendency
- **Measures of Variability (spread)**



Central Tendencies and Distributions

The representative value of a data set, generally the central value or the most occurring value that gives a general idea of the whole data set, is called **Measure of Central Tendency.**

Measures of Central Tendency
Some of the most commonly used measures of central tendency are:

- Mean
- Median
- Mode
- Skewness
- Kurtosis
- Variance
- Standard Deviation

Distribution Properties and Arithmetic

Distribution Properties

- **Mean (μ or M):** The average value of a dataset.
  - Calculated by summing all values and dividing by the number of values.
  - Example: For dataset {2, 4, 6}, Mean = (2+4+6)/3 = 4.
- **Median:** The middle value in a dataset when ordered from smallest to largest.
  - If the dataset has an even number of observations, the median is the average of the two middle numbers.
  - Example: For dataset {1, 3, 3, 6, 7, 8, 9}, Median = 6.
- **Mode:** The value that appears most frequently in a dataset.
  - A dataset can have more than one mode.
  - Example: For dataset {1, 2, 2, 3, 4}, Mode = 2.

Distribution Properties and Arithmetic

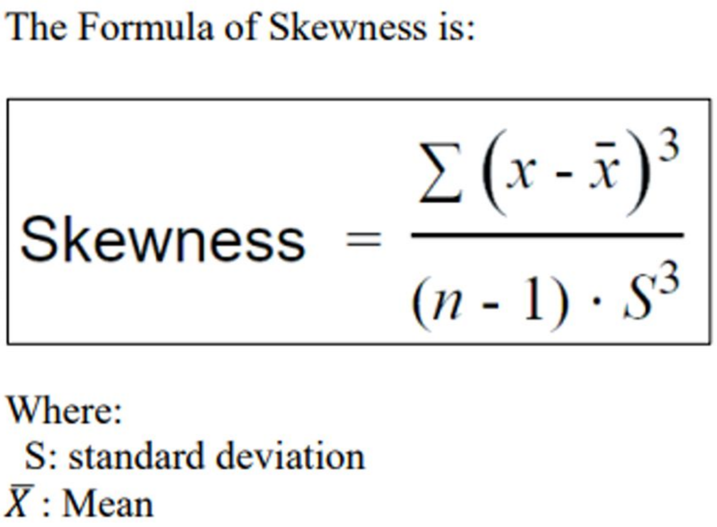
Distribution Properties

- **Variance (σ^2 or s^2):** A measure of how much values in a dataset deviate from the mean.
    - Calculated as the average of the squared differences from the mean.
    - Example: For dataset {2, 4, 6}, Variance = $[(2-4)^2 + (4-4)^2 + (6-4)^2]/3 = 2.67$.
- **Standard Deviation (σ or s):** The square root of the variance.
    - It gives an idea of the spread of values around the mean.
    - Example: For the above dataset, Standard Deviation = $\sqrt{2.67} \approx 1.63$.
- **Weighted Mean:** Mean where each value has a different weight.
    - Example: For values 3, 7, and 8 with weights 1, 2, and 3 respectively, Weighted Mean = (31 + 72 + 8*3)/(1+2+3) = 6.83.

Distribution Properties and Arithmetic

Distribution Properties

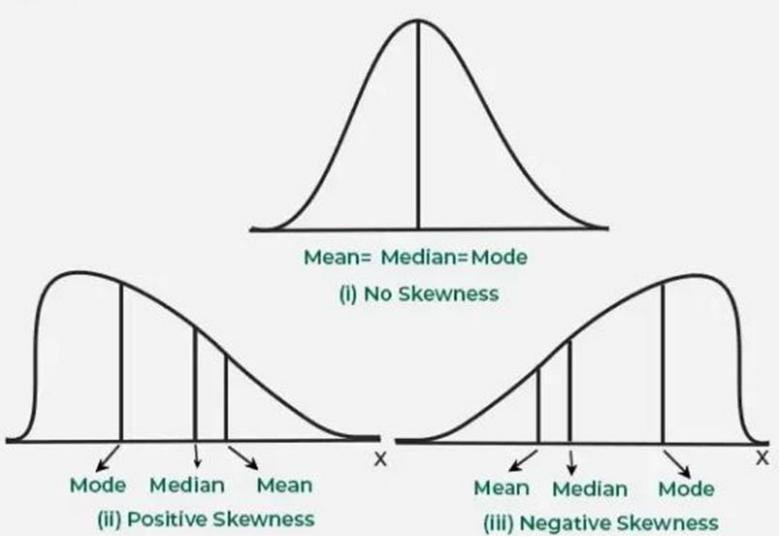
- **Skewness:** A measure of the asymmetry of the probability distribution of a real-valued random variable about its mean.



The Formula of Skewness is:

$$\boxed{\text{Skewness} = \frac{\sum (x - \bar{x})^3}{(n - 1) \cdot S^3}}$$

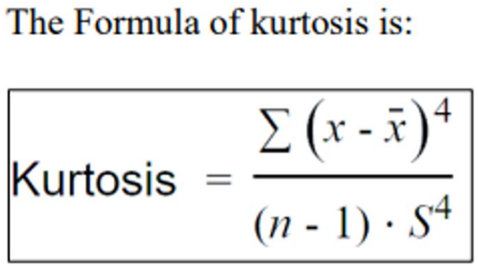
Where:
S: standard deviation
$\overline{X}$: Mean



Distribution Properties and Arithmetic

Distribution Properties

- **Kurtosis:** A measure of the **"tailedness"** of the probability distribution of a real-valued random variable.
- Kurtosis is a statistical measure that defines how heavily the tails of a distribution differ from the tails of a normal distribution. In other words, kurtosis identifies whether the tails of a given distribution contain extreme values.

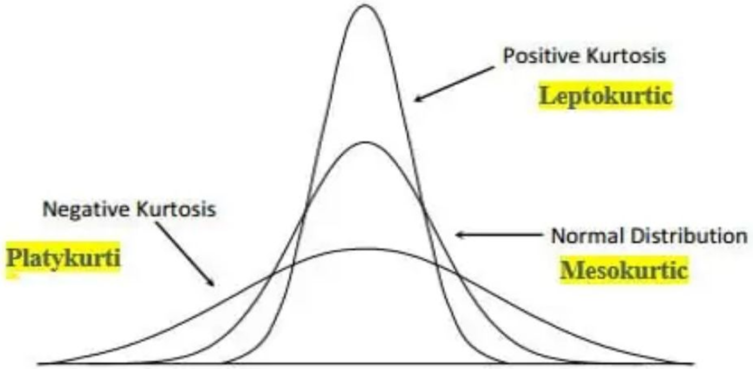


The Formula of kurtosis is:

$$\text{Kurtosis} = \frac{\sum\left(x - \bar{x}\right)^4}{(n - 1) \cdot S^4}$$

Where:

S: standard deviation $\bar{X}$: Mean



Inferential Statistics

Inferential statistics allow us to draw conclusions about a population based on a sample of data from that population.
Unlike descriptive statistics, which summarize data, inferential statistics make predictions and test hypotheses.

Key Concepts

1. Population vs. Sample:

- **Population:** The entire group you want to draw conclusions about (e.g., all B.Tech students).
- **Sample:** A subset of the population used to collect data (e.g., 100 B.Tech students from a particular university).

2. Parameter vs. Statistic:

- **Parameter:** A numerical characteristic of a population (e.g., population mean μ).
- **Statistic:** A numerical characteristic of a sample (e.g., sample mean $\bar{x}$).

3. Sampling Distribution:

- The probability distribution of a given statistic based on a random sample.
- For example, the sampling distribution of the sample mean approximates a normal distribution if the sample size is large enough (Central Limit Theorem).

Key Concepts

4. Confidence Intervals:

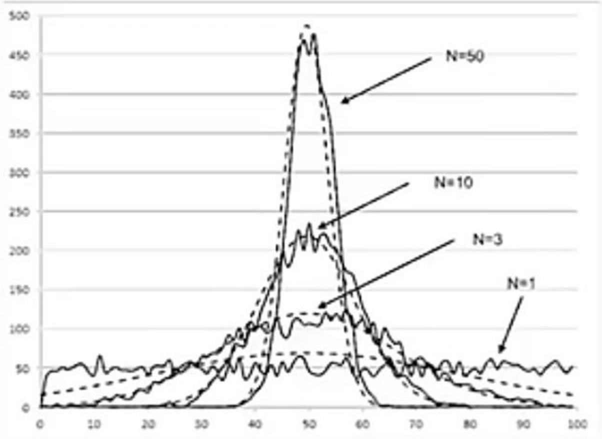
- A range of values that is likely to contain the population parameter.
- Example: A **95% confidence interval** for a mean might be **5 ± 1.96*SE**, where SE is the standard error.

5. Hypothesis Testing:

- A method to test if there is enough evidence to reject a null hypothesis (H0) in favor of an alternative hypothesis (H1).

CENTRAL LIMIT THEOREM

The ***Central Limit Theorem*** states that the sampling distribution of the mean of any independent, random variable will be normal or nearly normal, if the sample size is large enough



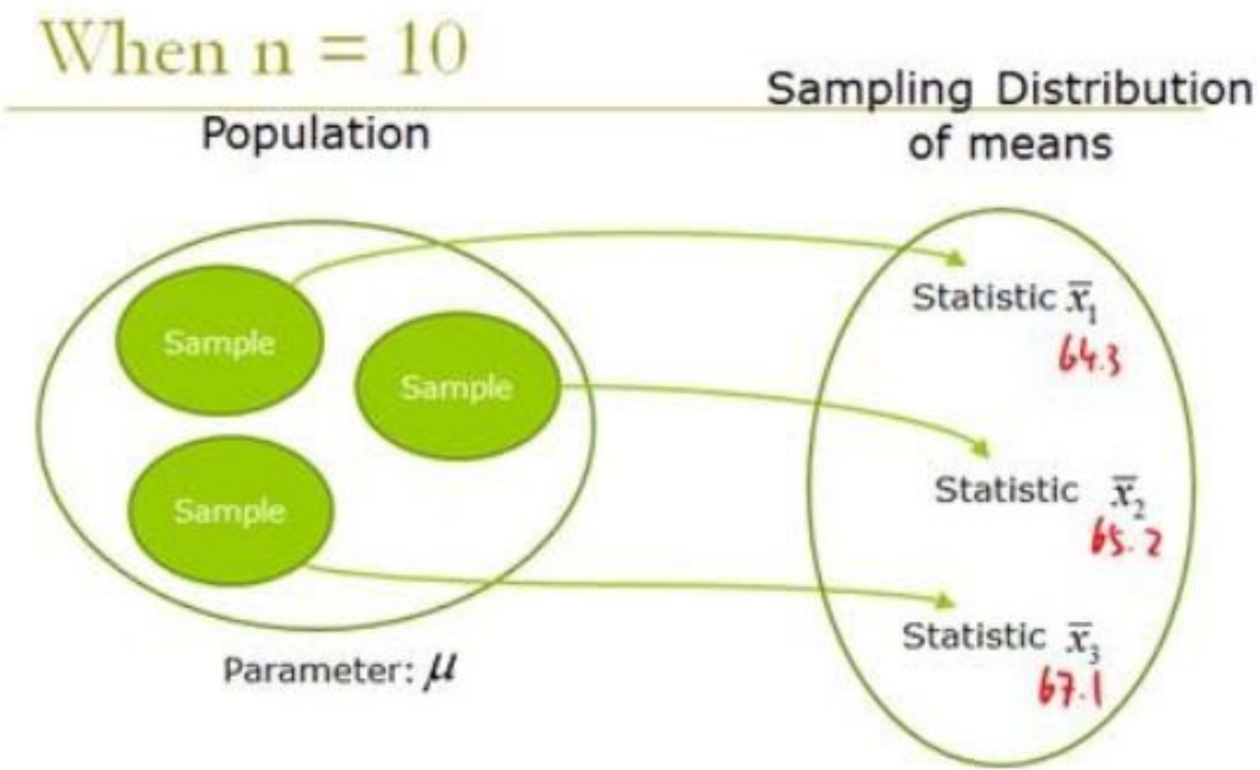
Central Limit Theorem with Example

- Consider that there are 15 sections in the science department of a university, and each section hosts around 100 students. Our task is to calculate the average weight of students in the science department. Sounds simple, right?
- The approach I get from aspiring data scientists is to simply calculate the average:
  - First, measure the weights of all the students in the science department.
  - Add all the weights.
  - Finally, divide the total sum of weights by the total number of students to get the average.
- But what if the size of the data is humongous? Does this approach make sense? Not really—measuring the weight of all the students will be a very tiresome and long process. So, what can we do instead? Let’s look at an alternate approach.

Let’s take the alternative approach...

- First, draw groups of students at random from the class. We will call this a sample. We’ll draw multiple samples, each consisting of 30 students.
- Now, calculate the individual mean of these samples.
- Then, calculate the mean of these sample means.
- This value will give us the approximate mean weight of the students in the science department.

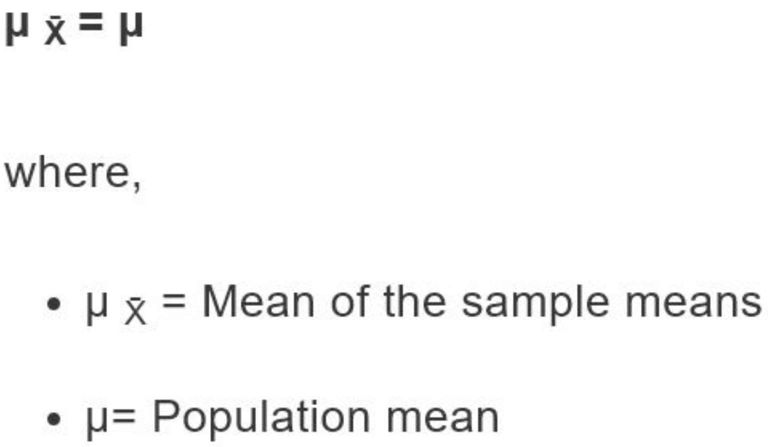
Central Limit Theorem



Assumptions Behind the Central Limit Theorem

- Before we dive into the implementation of the central limit theorem, it’s important to understand the assumptions behind this technique:
  - The **data must follow the randomization condition.** It must be sampled randomly
  - **Samples** should be **independent of each other**. One sample should not influence the other samples
  - **Sample size should be not more than 10% of the population** when sampling is done without replacement
  - The **sample size should be sufficiently large.** Now, how will we figure out how large this size should be? Well, it depends on the population.
  - When the **population is skewed or asymmetric,** the sample size should be large. If the population is symmetric, then we can draw small samples as well.
  - The **population’s distribution has a finite variance**. The central limit theorem doesn’t apply to distributions with infinite variance.

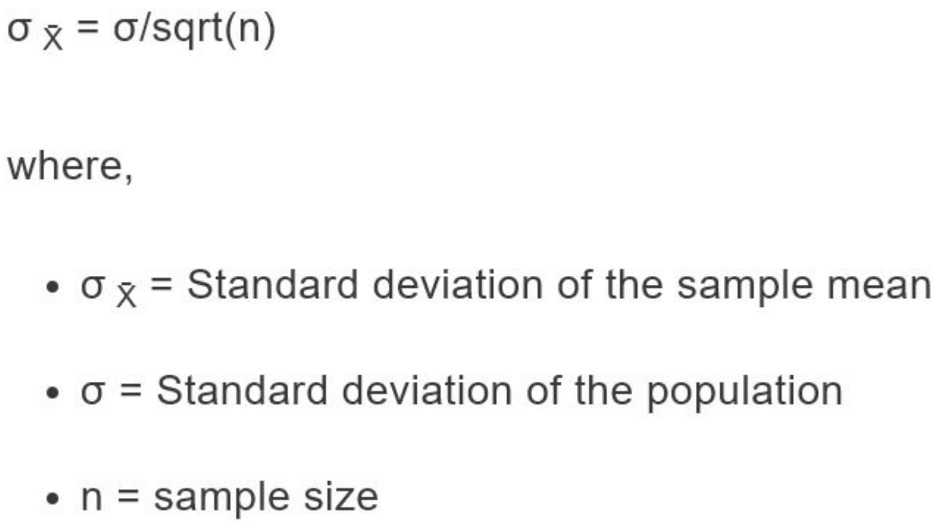
Central Limit Theorem



$\mu_{\bar{x}} = \mu$

where,

- $\mu_{\bar{x}}$ = Mean of the sample means
- μ= Population mean



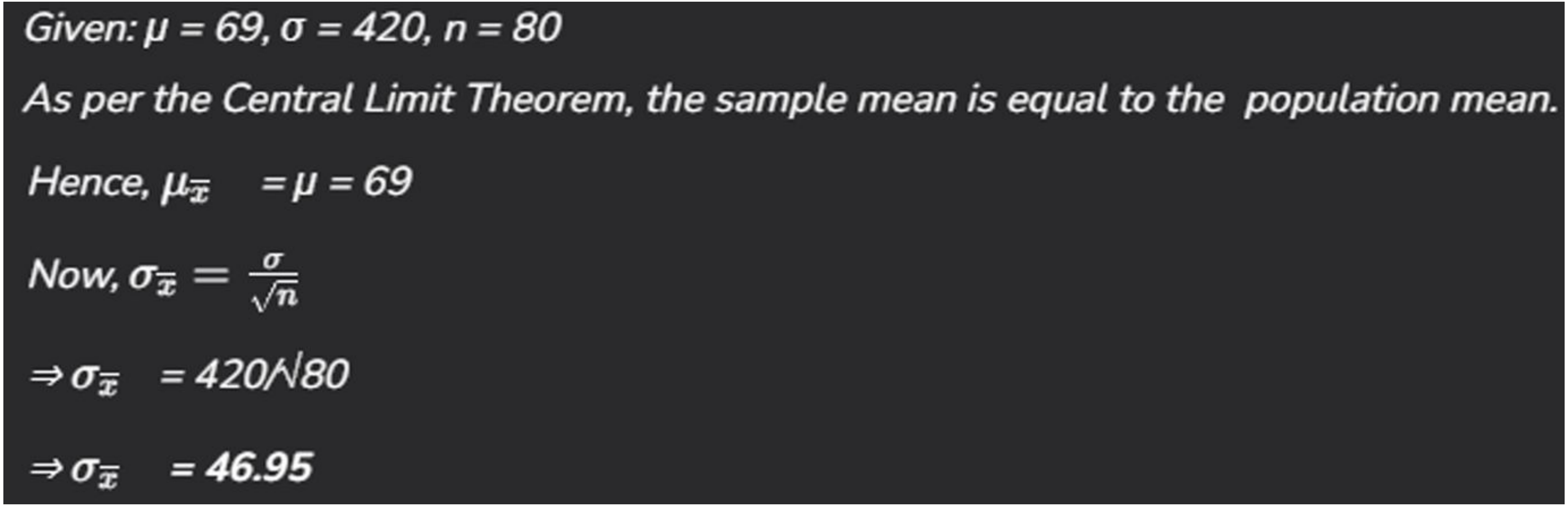
$\sigma_{\bar{x}}$ = σ/sqrt(n)

where,

- $\sigma_{\bar{x}}$ = Standard deviation of the sample mean
- σ = Standard deviation of the population
- n = sample size

Central Limit Theorem

- Example 1. A distribution has a mean of 69 and a standard deviation of 420. Find the mean and standard deviation if a sample of 80 is drawn from the distribution.



Given: $\mu = 69, \sigma = 420, n = 80$

As per the Central Limit Theorem, the sample mean is equal to the population mean.

Hence, $\mu_{\bar{x}} = \mu = 69$

Now, $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

$\Rightarrow \sigma_{\bar{x}} = 420/\sqrt{80}$

$\Rightarrow \sigma_{\bar{x}} = 46.95$

Central Limit Theorem

- **Example 2: The average weight of a water bottle is 30 kg, with a standard deviation of 1.5 kg. If a sample of 45 water bottles is selected at random from a consignment and their weights are measured, find the probability that the mean weight of the sample is less than 28 kg.**

Understand the Given Information:

- Population mean (p) = 30 kg
- Population standard deviation = 1.5 kg
- Sample size (n) = 45

We want to find the probability that the sample mean is less than 28 kg.

Central Limit Theorem

- **Example 2: The average weight of a water bottle is 30 kg, with a standard deviation of 1.5 kg. If a sample of 45 water bottles is selected at random from a consignment and their weights are measured, find the probability that the mean weight of the sample is less than 28 kg.**

The standard error of the mean is the standard deviation of the sampling distribution of the sample means. It's calculated as:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

$$\sigma_{\bar{x}} = \frac{1.5}{\sqrt{45}}$$

2. Z-score:

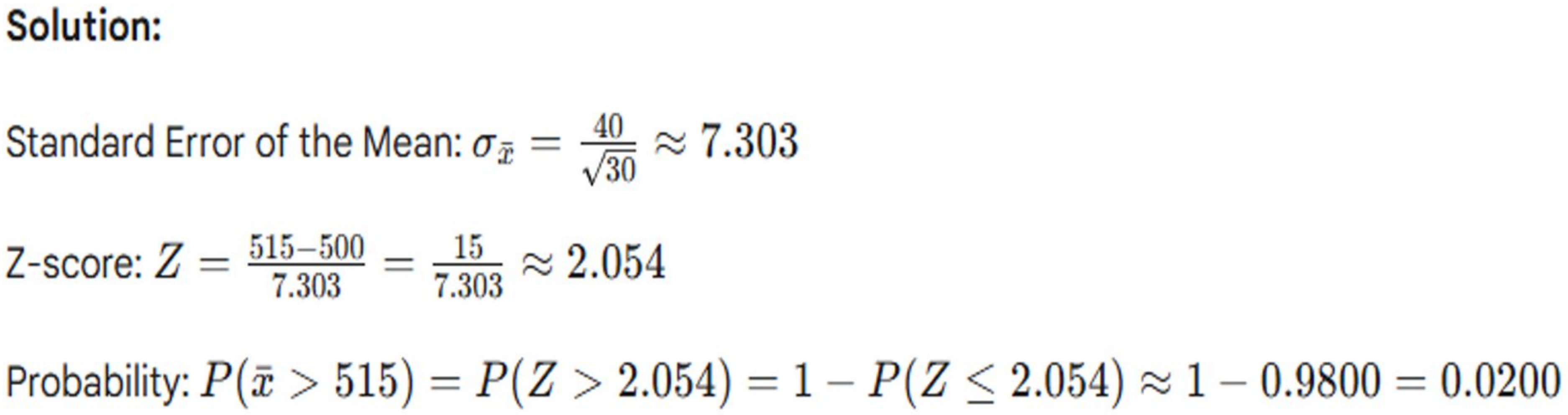
$$Z = \frac{\bar{x}-\mu}{\sigma_{\bar{x}}} = \frac{28-30}{0.2236} = \frac{-2}{0.2236} \approx -8.944$$

3. Probability:

$$P(\bar{x} < 28) = P(Z < -8.944) \approx 0.0000 \text{ (practically zero)}$$

Central Limit Theorem

- Example 3: The average daily sales of a coffee shop are 500 cups, with a standard deviation of 40 cups. If a random sample of 30 days is selected, what is the probability that the mean daily sales for this sample will be greater than 515 cups?



Solution:

Standard Error of the Mean: $\sigma_{\bar{x}} = \frac{40}{\sqrt{30}} \approx 7.303$

Z-score: $Z = \frac{515-500}{7.303} = \frac{15}{7.303} \approx 2.054$

Probability: $P(\bar{x} > 515) = P(Z > 2.054) = 1 - P(Z \leq 2.054) \approx 1 - 0.9800 = 0.0200$