STATISTICS

🔢 I. Types of Variables

1. Definition of Variable
2. Quantitative vs Qualitative Variables
3. Types of Quantitative Variables
    - Interval
    - Ratio
    - Key Differences

📊 II. Descriptive Statistics

4. Measures of Central Tendency
    - Mean
    - Median
    - Mode
5. Measures of Dispersion
    - Range
    - Variance
    - Standard Deviation
6. Percentiles and Quartiles
    - Percentile Rank
    - Quartile Definitions
7. Box Plot & Outlier Detection
    - IQR Method
    - Fence Calculations
    - Outlier Removal Example
8. Understanding Skewness
    - Positive vs Negative Skew

🎲 III. Probability & Random Variables

9. Basic Probability Concepts
    - Sample Space, Events
10. Conditional Probability & Independence
11. Random Variables

- Discrete
- Continuous

📈 IV. Probability Distributions

12. Overview of Distributions
13. Discrete Distributions

- Bernoulli
- Binomial
- Multinomial
- Poisson

14. Continuous Distributions

- Normal (Gaussian)
- Uniform
- Exponential

15. Comparison: Normal vs Uniform
16. Key Concepts

- PMF, PDF, CDF
- Mean, Variance, Standard Deviation

📐 V. Normal Distribution In-Depth

17. Properties of Gaussian Distribution
18. Standard Normal Distribution & Z-Score
19. Empirical Rule (68-95-99.7 Rule)
20. Applications
21. Q-Q Plot (Check for Normality)

🔍 **VI. Sampling Techniques**

22. What Makes a Good Sample?
23. Probability Sampling
    - Simple Random
    - Stratified
    - Systematic
    - Cluster
24. Non-Probability Sampling
    - Convenience Sampling

🧪 **VII. Inferential Statistics**

25. Central Limit Theorem (CLT)
26. Hypothesis Testing
    - Null & Alternative Hypotheses
    - Types of Hypothesis Tests (1-tailed, 2-tailed)
27. P-Value & Significance Level (α)
28. Confidence Intervals
29. Errors in Hypothesis Testing
    - Type I and Type II Errors
30. Statistical Tests
    - One-Sample T-Test
    - Two-Sample T-Test
    - Paired T-Test

🔧 **VIII. Estimation Techniques**

31. Maximum Likelihood Estimation (MLE)
    - Steps in MLE
    - Properties of MLE

📉 **IX. Handling Outliers**

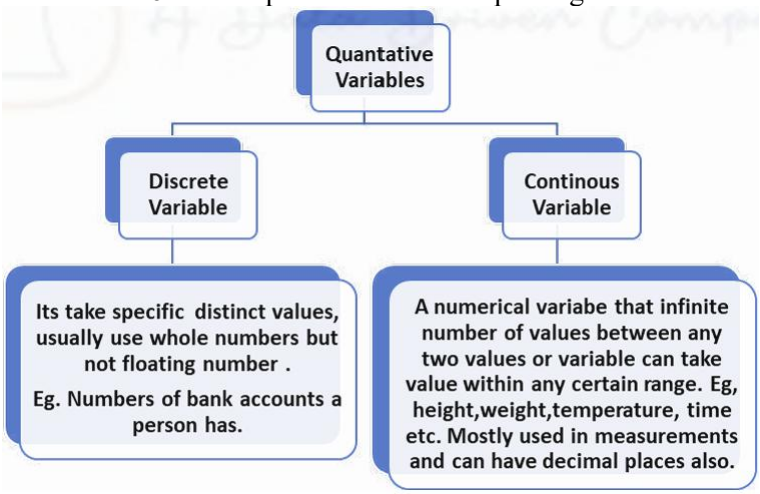
32. Negative & Positive Impacts of Outliers
33. Outlier Detection Methods
    - Box Plot, Z-Score, IQR
34. Outlier Treatment Techniques
    - Truncation, Winsorization, Transformation
    - Robust Models & Domain Expertise

🧩 **X. Handling Missing Data**

35. Understanding Missing Data Types
    - MCAR, MAR, MNAR
36. Imputation Methods
    - Mean/Median/Mode
    - KNN, Multiple Imputation, Model-Based
37. Alternatives to Imputation
    - Dropping Rows/Columns
    - Domain Expertise Involvement
38. Impact Assessment Post-Imputation

1. Variables

- **Variable**: A property that stores, manipulates, and retrieves data in a program.
- **Quantitative Variable**:
  - Involves numeric values (addition, subtraction, multiplication, division).
  - Example: Number of employees in a company.
- **Qualitative (Categorical) Variable**:
  - Represents categories or labels (no numerical meaning).
  - Example: Car colors in a parking lot.

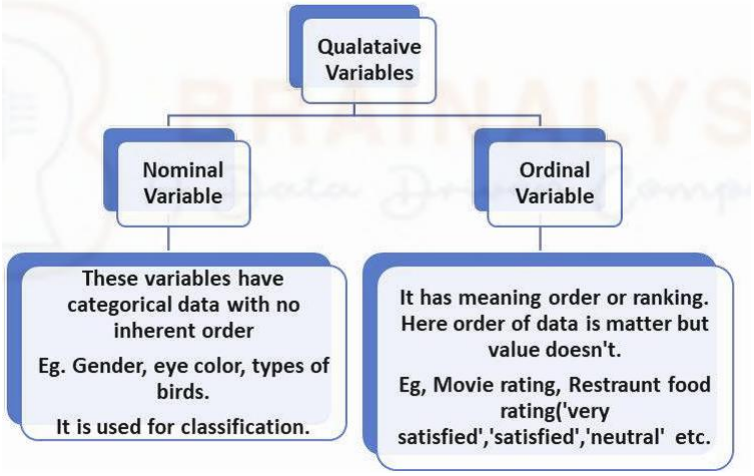


2. Types of Quantitative Variables

- **Interval Variable**:
  - Order and value matter but no true zero.
  - Example: Temperature in Fahrenheit (70°F → 80°F).
- **Ratio Variable**:
  - Has all properties of interval variables + a true zero point.
  - Example: Height (160 cm is twice as tall as 80 cm).

Key Difference Between Interval & Ratio:

- **Interval** lacks a true zero, so ratios are meaningless.
- **Ratio** has a true zero, allowing for meaningful comparisons.



---

3. Measures of Central Tendency

- Used to summarize the middle of a dataset.

A) Mean (Average)

- Formula: **Sum of all values ÷ Number of values**
- Example Dataset: {160, 165, 170, 175, 175, 180, 185, 185, 190, 190, 190}
- Calculation: Mean = 1965 / 11 ≈ 178.64
- **Python**:

```
import numpy as np
data = [160, 165, 170, 175, 175, 180, 185, 185, 190, 190, 190]
mean_value = np.mean(data)
```

- **SQL**:

SELECT AVG(values) FROM data;

B) Median (Middle Value in an Ordered Dataset)

- **If Odd**: The middle element.
- **If Even**: The average of the two middle elements.
- Example: {160, 165, 170, 175, 175, 180, 185, 185, 190, 190, 190}
  - Sorted: Median = **180** (5th value).
- **Python**:

median_value = np.median(data)

- **SQL**:

SELECT PERCENTILE_CONT(0.5) WITHIN GROUP (ORDER BY values) AS median FROM data;

C) Mode (Most Frequent Value)

- Example Dataset: {160, 165, 170, 175, 175, 180, 185, 185, 190, 190, 190}
  - Mode = **190** (most frequent value).
- **Python**:

```
from scipy import stats
mode_value = stats.mode(data).mode[0]
```

- **SQL**:

```
SELECT values
FROM data
GROUP BY values
ORDER BY COUNT(*) DESC
LIMIT 1;
```

Range: The diversity between the highest and lowest values. It offers an understanding of the spread of the data but may be influenced by anomalies.

Variance: How greatly the data points diverge from the average. It represents the mean of the squared discrepancies between each value and the average.

Population Variance :

$$\sigma^2 = \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{N}$$

} Degree of Freedom

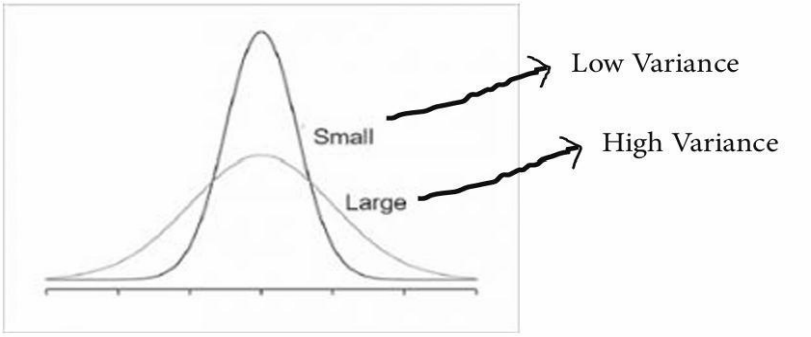
Sample Variance :

$$s^2 = \sum_{i=1}^{n} \frac{(x_i-\bar{x})^2}{n-1}$$

Sample

$$S = \sqrt{\frac{\Sigma(x_i-\bar{x})^2}{n-1}}$$

$\bar{x}$ - Sample Average
x_i - Individual Population Value
n - Total Number of Sample



Why use "n - 1" within the denominator instead of "n" within the equation (dividing by using "n - 1" as opposed to "n")?

1. Degrees of Liberty (df): Degrees of freedom resemble the amount of jigsaw portions we can rearrange without absolutely decoding the complete photo. In records, they demonstrate the power of our calculations while making sure fairness.

2. Sample Variability: Picture this: we attempt to ascertain how broadly dispersed facts is inside a small cluster (sample). The sample mean aids in estimating how facts behaves inside the large cluster (population). However, this proves hard due to the fact the pattern does not encompass the whole populace.

3. The Predicament with "n": Opting solely for "n" (the records points general) as the denominator in the variance system could yield a skewed outcome. This discrepancy arises due to the fact the sample imply diverges from the population mean, complicating the calculations.

4. Introduction of Bessel's Correction ("n - 1"): To rectify this issue, Bessel's correction is delivered. Rather than using "n," we replace it with "n - 1" inside the denominator. This correction acknowledges that we're extrapolating from a sample, no longer the whole populace, thereby allowing more leeway for the data to vary.

5. Significance of "n - 1": Incorporating "n - 1" in preference to "n" complements the accuracy of our projection concerning the overall populace's dispersal. This approach averts the understatement of variability within the population primarily based on the pattern, particularly crucial while handling minute samples.

Conclusion: So, Bessel's correction is a tweak that makes sure where variance calculations are better when we're dealing with samples. It's like adding a little extra flexibility to were calculations to match the real world better. Bessel's correction addresses the issue of underestimating variability in sample-based calculations and ensures that the estimated variance is a better representation of the population variance

Key take aways: Spread is low means the elements present in the central region is more.
More variance: Data is more spread.
Variance = Spread = Dispersion = Is the extent to which distribution is stretched or squeezed

Standard Deviation: The square root of variance. It's a commonly used measure of spread, indicating how much data tends to deviate from the mean. It shows how far the elements are from mean

$$\sigma = \sqrt{variance}$$

Calculate the standard deviation of the values 45, 35, 42, 49, 39, and 34. Give your answer to 3 decimal places.

Mean = $\frac{122}{3}$

Standard Deviation = $\sqrt{\text{Variance}}$ (σ)

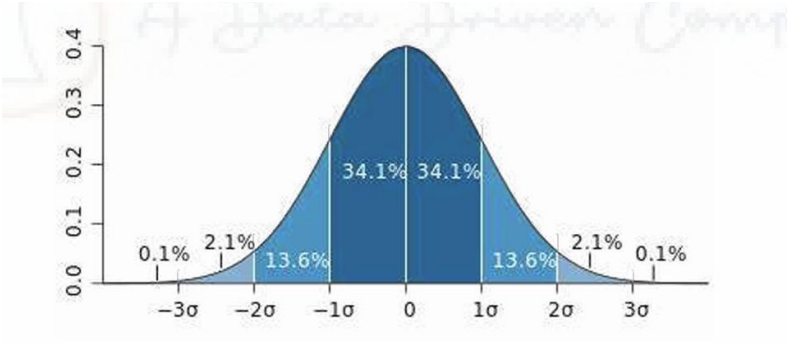
Variance = $\frac{254}{9}$

↑ average of the squared differences from the mean

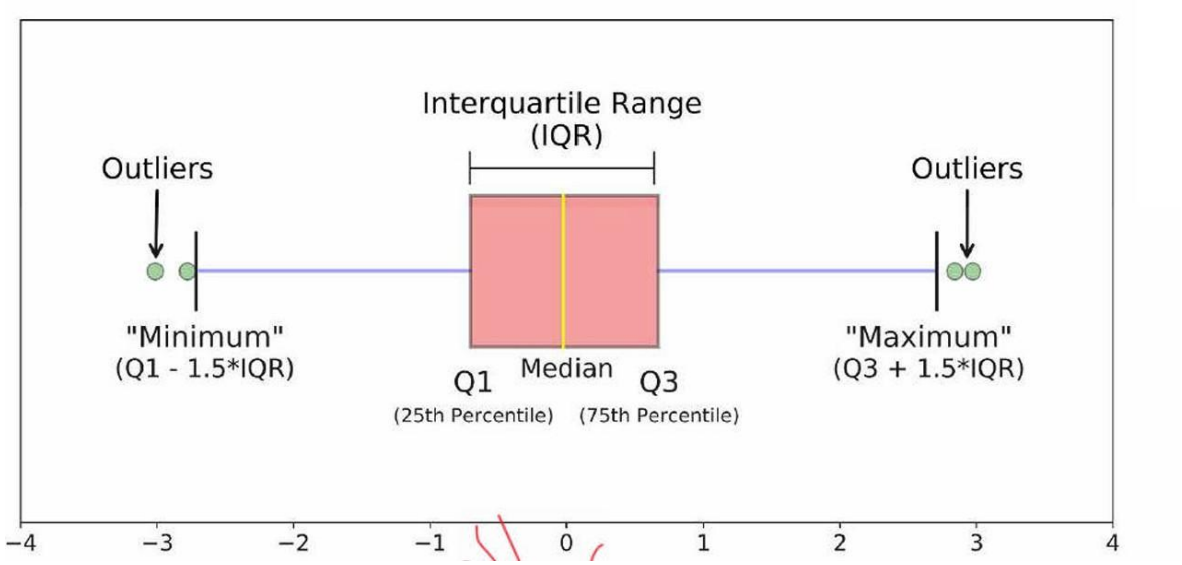
$$\text{Variance} = \frac{\left(\frac{122}{3}-45\right)^2+\left(\frac{122}{3}-35\right)^2+\left(\frac{122}{3}-42\right)^2+\left(\frac{122}{3}-49\right)^2+\left(\frac{122}{3}-39\right)^2+\left(\frac{122}{3}-34\right)^2}{6}$$

$$\text{Standard Deviation} = \sqrt{\frac{254}{9}} = 5.312345\ldots$$

5.312




```python
data = np.array([12, 15, 20, 22, 18, 30, 25, 28, 17, 19])
# Range
data_range = np.max(data) - np.min(data)
# Variance (Sample variance: ddof=1)
variance = np.var(data, ddof=1)
# Standard Deviation (Sample std dev: ddof=1)
std_dev = np.std(data, ddof=1)
```



Percentiles and Quartiles (For Locating Data Points)

- **Percentiles:**
  - Divide data into 100 sections.
  - A percentile represents the value below which a certain percentage of observations fall.
  - Example:
    - 75th percentile → 75% of individuals have values below it.
    - **Finding a Percentile Rank:**
      - Example Dataset: {2, 2, 3, 4, 5, 5, 5, 6, 7, 8, 8, 8, 8, 8, 9, 9, 10, 11, 11, 12}
      - Percentile rank calculation:
      - First decide total number of values in dataset (n=20)
      - Pin point where the cost 10 stands in the ordered set (on the 17th role)
      - Calculation = (Position of 10 / Total values) * 100
      - (17 / 20) * 100 = 85
      - So, **10 falls at the 85th percentile**.

- **Quartiles (Divide Data into Four Equal Parts)**
  - **Minimum:** Smallest value in the dataset.
  - **First Quartile (Q1):** 25% of data falls below this.
  - **Median (Q2):** 50% of data falls below this (middle value).
  - **Third Quartile (Q3):** 75% of data falls below this.
  - **Maximum:** Largest value in the dataset.

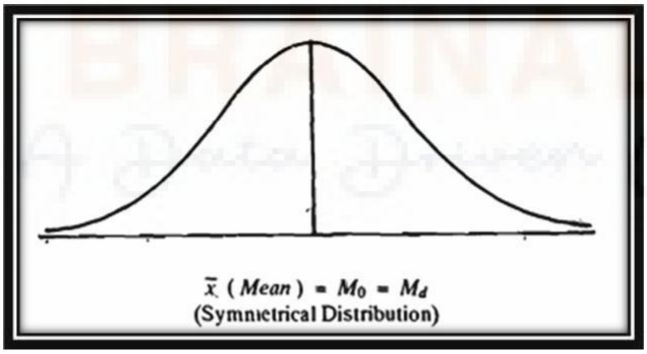
Box Plot (For Outlier Removal)

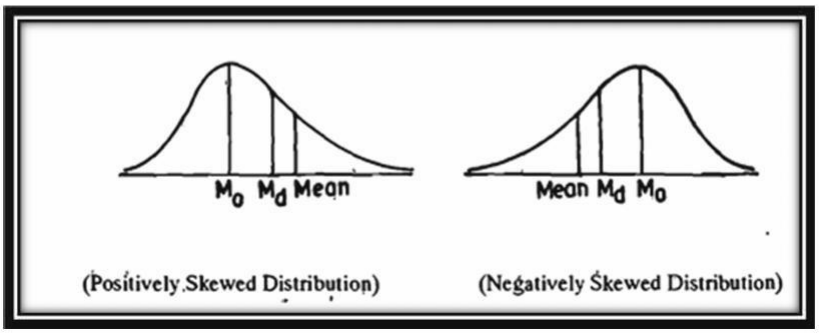
6. **Original Dataset (With Outliers):**
   {1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 27}
7. **Find Quartiles:**
   - Q1 (5th value) = **3**
   - Q3 (15th value) = **8**
8. **Interquartile Range (IQR):**
   - IQR = **Q3 - Q1 = 8 - 3 = 5**
9. **Calculate Fences for Outlier Detection:**
   - Lower Fence = **Q1 - 1.5 × IQR = 3 - 1.5 × 5 = -4.5**
   - Upper Fence = **Q3 + 1.5 × IQR = 8 + 1.5 × 5 = 15.5**
10. **Remove Outliers (Values Outside Fences):**
    - Remove values **< -4.5 or > 15.5** → Remove **27**
    - Remaining dataset: {1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9}
11. **Recalculate Quartiles for Cleaned Data:**
    - **Minimum:** 1
    - **Q1:** 3
    - **Median (Q2):** 5
    - **Q3:** 8
    - **Maximum:** 9

```
# Calculate Q1 (25th percentile) and Q3 (75th percentile)
q1 = np.percentile(data, 25)
q3 = np.percentile(data, 75)
# Interquartile Range (IQR)
iqr = q3 - q1
# Outlier boundaries
lower_bound = q1 - 1.5 * iqr
upper_bound = q3 + 1.5 * iqr
# Identify outliers
outliers = data[(data < lower_bound) | (data > upper_bound)]
```

Understanding Skewness

- **Skewness:**
  - Refers to the **absence of symmetry** in a dataset.
  - Indicates how values in a dataset are distributed.
  - A distribution is **skewed** if:
    - The **mean, median, and mode** are not aligned.
    - The **median is unequally spaced** from the quartiles.
    - The **graph is asymmetrical**, leaning more towards one side.
- **Types of Skewness:**
  - **Positive Skewness (Right Skewed):**
    - The **tail extends to the right**.
    - **Mean > Median > Mode**.
  - **Negative Skewness (Left Skewed):**
    - The **tail extends to the left**.
    - **Mean < Median < Mode**.





Kurtosis

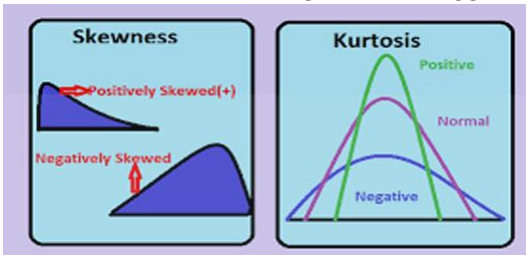
• Kurtosis measures how much a dataset's distribution deviates from a regular distribution, specifically in phases of the tails' heaviness.

• It measures things like central tendency, dispersion, and skewness. Kurtosis provides vital insights into distribution of data.

• Kurtosis describes the shape of the distribution's tails and its weight.

• Higher kurtosis shows heavier tails suggesting more frequent extreme values.

• Lower kurtosis shows lighter tails suggesting lesser extreme values.



Types of Sampling

What is a good sample?

A Good sample should represent the entire population and each member should have equal chance to become a part of sample.

Types of Sampling – PROBABLITY SAMPLING AND NON-PROBABLITY SAMPLING

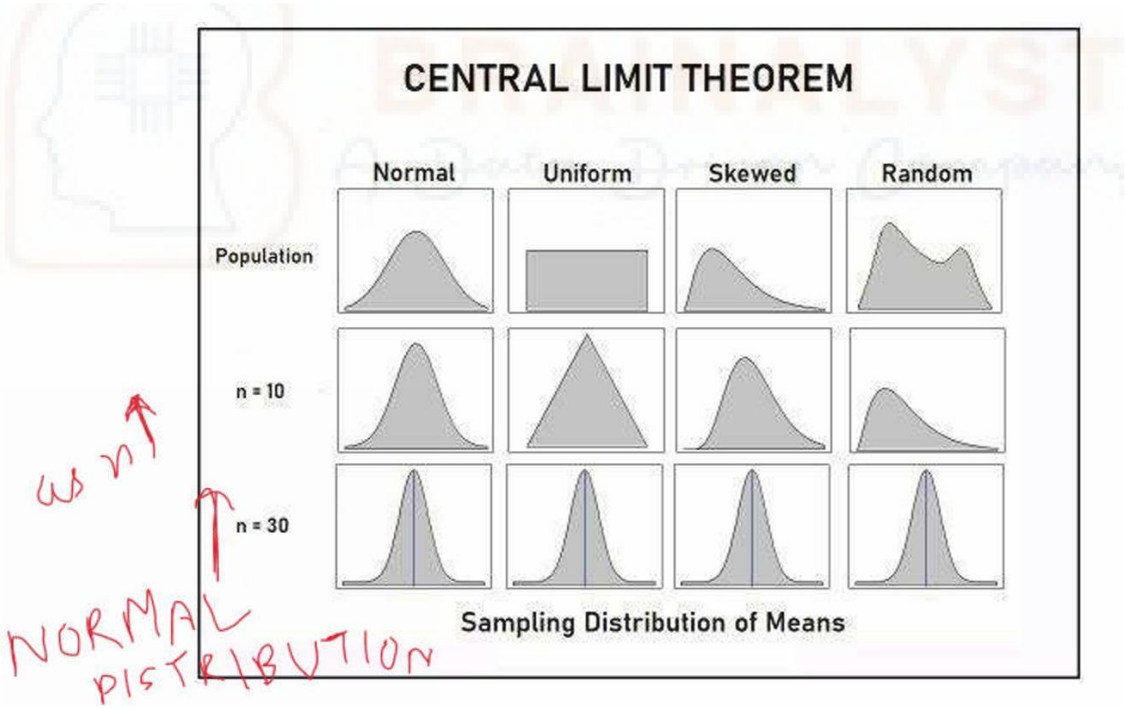
➔ PROBABLITY SAMPLING

- **Simple Random Sampling:**
  - Every member of the population has an **equal chance** of being selected.
  - Ensures unbiased representation.
- **Stratified Sampling:**
  - The population is **divided into separate, non-overlapping groups** (strata).
  - A sample is taken **from each group**, ensuring better representation of subgroups.
- **Systematic Sampling:**
  - Members are **selected at a fixed interval (nth interval)** from the population.
  - Example: Picking every **"nth" item** from a list

- **Clustered Sampling**
  o The population is divided into groups or clusters, usually based on geography or other natural groupings.
  o Then, entire clusters are randomly selected, and either all members of selected clusters or a random sample from them are surveyed.
  o Example: Randomly selecting a few schools from a district and surveying all students in those selected schools.

- **Convenience Sampling:**
  - Members are **chosen based on availability and ease of access**.
  - Often used when **time or resources are limited**, but may lead to bias.

CENTRAL LIMIT THEOREM



Definition:

- The **Central Limit Theorem (CLT)** states that when a **large enough sample** is taken from a population, the sample **means will form a normal distribution**, regardless of the original population's distribution.
- As the **sample size increases**, the sample mean **approaches the population mean**, and **variability decreases**.

Why is CLT Important?

- Helps analyze **non-normal data distributions**.
- Enables **accurate estimations** of a population's characteristics using **smaller samples**.
- Useful because **real-world data is often unstructured and irregular**.
- Allows researchers to **work with samples instead of measuring the entire population**.

Key Takeaways:

- With **larger sample sizes**, the sample means **closely resemble a normal distribution**.
- **Simplifies statistical analysis** by allowing the use of normal distribution techniques.

Standard Error (S.E.):
Definition:
Standard Error measures how much a statistic's outcome (like a mean or proportion) can vary across

different samples from the same population. It quantifies the **uncertainty** or **"wiggle room"** in a sample's estimate of the population parameter.

Why It's Important:

- **Measure of Confidence:** SE helps estimate how far the sample statistic is likely to be from the true population parameter.
- **Reliability Check:** It shows whether the results are consistent and reliable across multiple samples or whether they may fluctuate significantly.

Common Formulas:

1. **Standard Error of the Mean ($SE_{\bar{x}}$):**

$$SE = \frac{\sigma}{\sqrt{n}}$$

- σ = Population standard deviation
- n = Sample size

2. **Standard Error of a Proportion (SE_p):**

$$SE = \sqrt{\frac{p(1-p)}{n}}$$

- p = Sample proportion
- n = Sample size

3. **Standard Error of the Difference Between Two Means:**

$$SE = \sqrt{\left(\frac{\sigma_1^2}{n_1}\right) + \left(\frac{\sigma_2^2}{n_2}\right)}$$

- σ_1 and σ_2 = Standard deviations of the two populations
- n_1 and n_2 = Corresponding sample sizes

Where It's Used:

- **Research & Surveys:** To assess how well a sample represents the population.
- **Comparative Studies:** To test if observed differences (e.g., between groups or treatments) are statistically meaningful or due to random variation.
- **Reports & Presentations:** SE is reported alongside estimates to indicate the level of certainty or margin of error.

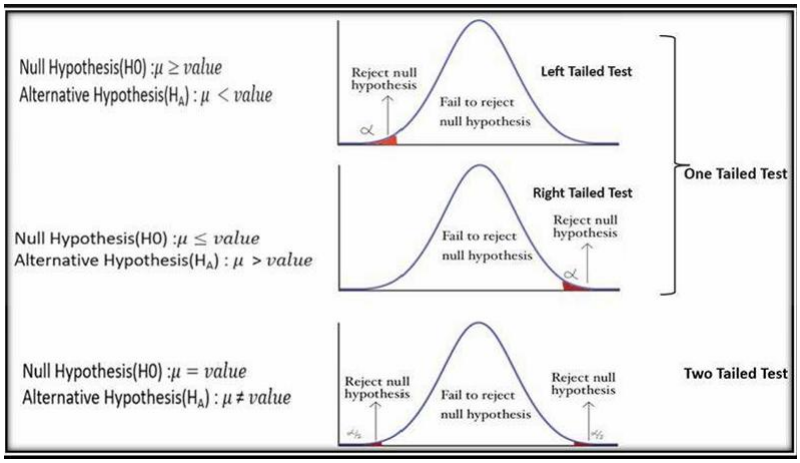
Key Takeaway:

The **smaller** the standard error, the **more reliable** the sample statistic. As sample size increases, standard error decreases, enhancing the accuracy of population estimate.

Hypothesis Testing: Comparing Two Organizations

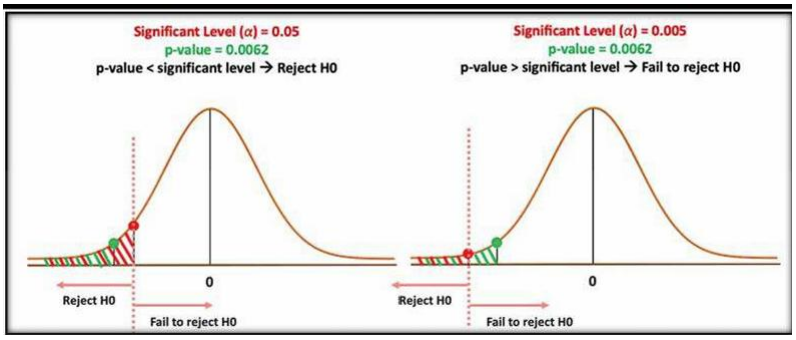
- Hypothesis testing is like being a detective—gathering evidence (statistics) to support or refute a claim.
- **Key Components:**
  - **Null Hypothesis (H0):** No significant difference or effect.
    - Example: A new drug has no effect on sleep.
  - **Alternative Hypothesis (Ha or H1):** There is a significant difference or effect.
    - Example: A new drug improves sleep quality.

- **Types of Alternative Hypotheses:**
  - **Two-Tailed:** The mean is different (not equal) from a given value.
  - **Right-Tailed:** The mean is greater than the given value.
  - **Left-Tailed:** The mean is less than the given value.



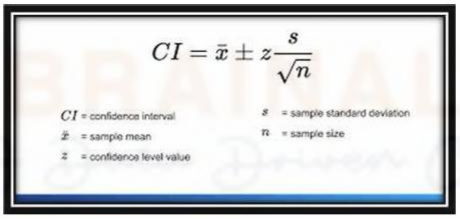
P-Value and Significance Level (α)

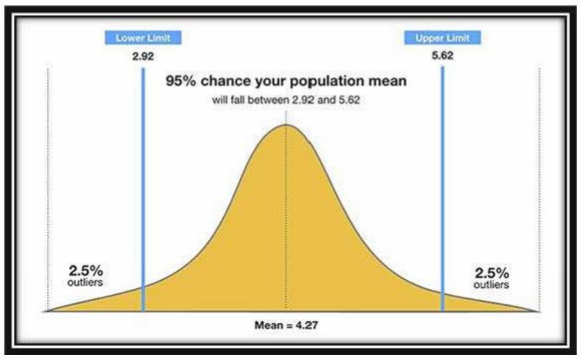
- **P-Value:** Probability of observing sample results under the null hypothesis.
  - Small p-value (e.g., < 0.05) → Reject the null hypothesis.
  - Large p-value → Fail to reject the null hypothesis.
  - Ranges between **0 and 1** (0 = impossible under H0, 1 = very likely).
- **Significance Level (α):** Threshold for rejecting H0.
  - If **p-value ≤ α**, reject H0.
  - If **p-value > α**, fail to reject H0.



Confidence Intervals

- Range of values where the true population parameter is likely to fall.
- Example: "We're 95% confident that the average score is between 80 and 90.





Errors in Hypothesis Testing

- **Type I Error (False Positive):** Rejecting a true null hypothesis.
  - Example: A test wrongly detects HIV in a healthy person.
- **Type II Error (False Negative):** Failing to reject a false null hypothesis.
  - Example: A test fails to detect HIV in an infected person.

- **Preference:**
  - In medical tests, Type I errors (false positives) are preferred to avoid missing real cases.
  - In scientific research, Type I errors are considered more serious.

Statistical Tests for Hypothesis Testing

- **T-Test:** Compares means between groups.
  - **Assumptions:** Normal distribution, equal variances.
- **Types of T-Tests:**
  - **One-Sample T-Test:**
    - Compares sample mean with a known value.
    - **Example:** Testing if the population mean IQ (VIQ measure) is 0.

```
In [10]: from scipy import stats

         # Sample data
         sample_data = [23, 25, 22, 26, 28, 24, 27, 26, 29, 30]

         # Hypothesized population mean
         population_mean = 28

         t_stat, p_value = stats.ttest_1samp(sample_data, population_mean)
         if p_value < 0.05:
             print("Reject null hypothesis: Sample mean is significantly different from population mean.")
         else:
             print("Fail to reject null hypothesis: No significant difference.")

         Reject null hypothesis: Sample mean is significantly different from population mean.
```

  - **Two-Sample T-Test:**
    - Compares means of two independent samples.
    - Example: Testing if male and female VIQ means are significantly different.

```
In [11]: from scipy import stats

         # Sample data for two groups
         group1 = [15, 18, 20, 22, 23]
         group2 = [25, 27, 28, 30, 32]

         t_stat, p_value = stats.ttest_ind(group1, group2)
         if p_value < 0.05:
             print("Reject null hypothesis: There is a significant difference between the group means.")
         else:
             print("Fail to reject null hypothesis: No significant difference.")

         Reject null hypothesis: There is a significant difference between the group means.
```

  - **Paired T-Test:**
    - Used for repeated measurements on the same individuals (e.g., before and after tests).
    - Example: Comparing FSIQ and PIQ scores for the same person.

Interpret the Results:

- If the p-value is less than the chosen importance level (α), you could reject the null hypothesis and finish that there may be a widespread distinction between the manner.
- If the p-value is extra than or equal to the significance level, you fail to reject the null speculation, indicating that there is no great difference between the approaches.

Keep in mind that the translation of the p-cost relies upon the chosen importance stage. A smaller p-value shows stronger evidence for null speculation.

Chi-Square Test

• **Chi-Square Test**: Shows is there is a strong link between two data types. Assesses the association between categorical variables.
o **Assumptions**: Expected frequency ≥ 5 in each cell, independent observations.
• **Types of Chi-Square Tests**:
o **Chi-Square Test for Independence**:
♣ Tests if two categorical variables are independent.
o **Chi-Square Goodness-of-Fit Test**:

♣ Compares observed vs. expected frequencies in one categorical variable. (if sample distribution fits the population distribution)

ANOVA (Analysis of Variance)

• **ANOVA**: Compares means across multiple groups.
o **Assumptions**: Normal distribution, homogeneity of variance, independent observations.
• **Types of ANOVA**:
o **One-Way ANOVA**:
♣ Compares means across three or more independent groups.
o **Two-Way ANOVA**:
♣ Examines the effect of two independent variables on a dependent variable.
o **Repeated Measures ANOVA**:
♣ Compares means within the same group at different time points.

```
import numpy as np
from scipy.stats import chi2_contingency
# Step 1: Create the observed frequency table
# Rows: Gender (Male, Female)
# Columns: Preference (Tea, Coffee)
observed = np.array([
    [30, 10],  # Male
    [20, 40]   # Female
])
# Step 2: Apply the chi-square test
chi2_stat, p_val, dof, expected = chi2_contingency(observed)

# Step 3: Interpret the result
if p_val < 0.05:
    print("Result: Dependent (reject null hypothesis)")
else:
    print("Result: Independent (fail to reject null hypothesis)")
```

Mann-Whitney U Test

• **Mann-Whitney U Test**: Non-parametric test comparing two independent groups when assumptions of a t-test aren't met.
o **Assumptions**: Ordinal or continuous data, independent groups, similar distribution shapes.
• **Usage**:
o **Alternative to the Independent T-Test** when normality assumption is violated.

```
import numpy as np
from scipy.stats import mannwhitneyu
# Scores of two groups (independent samples)
group_a = [88, 92, 85, 91, 87]
group_b = [75, 78, 80, 70, 73]
# Perform Mann-Whitney U Test
stat, p = mannwhitneyu(group_a, group_b, alternative='two-sided')
# Print the result
print("Mann-Whitney U statistic:", stat)
print("p-value:", p)
# Interpretation
if p < 0.05:
    print("Result: Significant difference between groups (reject H₀)")
else:
    print("Result: No significant difference (fail to reject H₀)")
```

Z-Test

• **Z-Test**: Compares means when sample size is large (n ≥ 30).
o **Assumptions**: Normal distribution, known population variance, independent observations.
• **Types of Z-Tests**:
o **One-Sample Z-Test**:
♣ Compares the mean of a single group to a known population mean.
o **Two-Sample Z-Test**:
♣ Compares the means of two independent groups.
o **Z-Test for Proportions**:
♣ Compares proportions between two groups.

```python
import numpy as np
from statsmodels.stats.weightstats import ztest

# Sample data
sample = np.array([72, 71, 69, 70, 74, 68, 72, 73, 69, 71])

# Perform one-sample Z-test (H0: sample mean = 70)
z_stat, p_val = ztest(sample, value=70)

print("Z-statistic:", z_stat)
print("p-value:", p_val)

# Interpretation
if p_val < 0.05:
    print("Result: Significant difference (reject H₀)")
else:
    print("Result: No significant difference (fail to reject H₀)")
```

Choosing Statistical Tests for Comparing Two Groups
• Key Factors:
o Data Type: Continuous vs. Categorical.
o Sample Size: Small vs. Large.
o Distribution: Normal vs. Non-Normal.
• Test Selection:
o T-Test: Compare means of two independent or paired groups (normal distribution).
o Mann-Whitney U Test: Non-parametric alternative to T-Test.
o Chi-Square Test: Compare categorical variables.
o Z-Test: Compare means when sample size is large (n ≥ 30).
o ANOVA: Compare means of three or more groups.

Data Type	Group Type	Normality	Test
Continuous	Independent	Normal	T-Test
Continuous	Independent	Non-Normal	Mann-Whitney U
Continuous	Paired	Normal	Paired T-Test
Continuous	Paired	Non-Normal	Wilcoxon Signed-Rank
Categorical	-	Large Sample	Chi-Square
Categorical	-	Small Sample	Fisher's Exact
Proportions	Independent	-	Z-Test for Proportions

Question Type	**Test**
Is the mean different from a known value?	One-sample t-test
Are two means different?	Independent or paired t-test
Are medians different?	Mann-Whitney U / Wilcoxon
Are more than two means different?	ANOVA / Kruskal-Wallis
Are proportions different?	Z-test for proportions / Chi-square
Are variances different?	F-test / Levene's test
Are variables correlated?	Pearson / Spearman
Is a factor predictive of outcome?	Regression (Linear/Logistic)

📊 Hypothesis Testing Reference Table

Test Name	**When to Use**	**Test Statistic (t/z/F)**	**p-value**	**Conclusion**
Z-test	Proportion comparison (large n, known σ, binary metric like conversion rate)	$z > 1.96$ (for 95% CI)	$p < 0.05$	Reject H_0 → Significant difference
t-test	Comparing two means (small n, unknown σ, normal-ish distribution)	$t > 2$ (approx for df>30)	$p < 0.05$	Reject H_0 → Means are significantly different
Paired t-test	Same users tested before & after a change (e.g., CTR before/after redesign)	$t > 2$	$p < 0.05$	Reject H_0 → Treatment changed the metric
ANOVA (F-test)	Comparing more than 2 groups (e.g., 3 versions of a feature)	$F > 3$ (depends on df)	$p < 0.05$	Reject H_0 → At least one group differs
Chi-square test	Categorical variables (e.g., click vs no-click across regions)	$\chi^2 >$ critical value	$p < 0.05$	Reject H_0 → Distribution is not due to chance
Mann-Whitney U	Non-parametric comparison of two independent groups (skewed data)	$U <$ critical value	$p < 0.05$	Reject H_0 → Groups are significantly different
Wilcoxon	Non-parametric for paired data	$W <$ critical	$p <$	Reject H_0 → Change is

Signed-Rank	(before/after, small n)	value	0.05	significant

🧪 **Understanding the Components:**

- **p-value < 0.05** → Evidence **against null hypothesis** (i.e., significant difference)
- **p-value ≥ 0.05** → **Fail to reject** H_0 (i.e., difference not statistically significant)
- **Z/t/F statistic** must exceed **critical values** to be significant (based on confidence level)
  - For **95% CI**, Z critical = ±1.96
  - For **99% CI**, Z critical = ±2.58

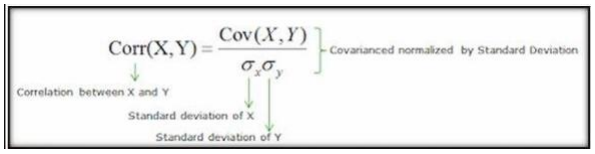
Example (Z-test in Swiggy A/B Test):

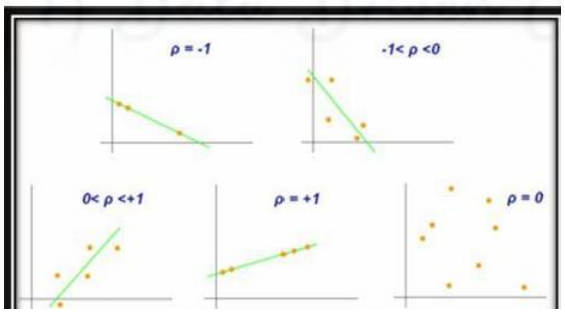
Group	Users	Conversion Rate
A	10,000	32%
B	10,000	35%

- **z = 2.57**, **p = 0.01** → $p < 0.05$
  → ✅ **Conclusion**: Version B has a statistically significant improvement

Correlation

- Correlation: Measures the strength and direction of a relationship between two variables.
  o Assumptions: Linearity, continuous variables, no extreme outliers.
  • Types of Correlation:
  o Pearson Correlation: Measures linear relationships. The value stages in between -1 to 1.





o Spearman Correlation: Measures rank-based relationships. It focuses on order rather than exact values
o Kendall's Tau: Measures ordinal association.
• Interpretation:
o +1: Strong positive correlation.
o -1: Strong negative correlation.
o 0: No correlation.

Covariance	Correlation
Indicates the direction of the linear relationship between variables	Indicates both the strength and direction of the linear relationship between two variables
Covariance values are not standard	Correlation values are standardized
Positive number being positive relationship and negative number being negative relationship	1 being strong positive correlation, -1 being strong negative correlation
Value between positive infinity to negative infinity	Value is strictly between -1 to 1

```python
import numpy as np
from scipy.stats import pearsonr, spearmanr, kendalltau

# Sample data
x = [10, 20, 30, 40, 50]
y = [12, 24, 33, 47, 55]

# Pearson Correlation
pearson_corr, pearson_p = pearsonr(x, y)
# Spearman Correlation
spearman_corr, spearman_p = spearmanr(x, y)
# Kendall Tau Correlation
kendall_corr, kendall_p = kendalltau(x, y)

# Print all results
print("=== Correlation Coefficients ===")
print(f"Pearson : {pearson_corr:.4f} (p={pearson_p:.4f})")
print(f"Spearman: {spearman_corr:.4f} (p={spearman_p:.4f})")
print(f"Kendall : {kendall_corr:.4f} (p={kendall_p:.4f})")

# Optional: Quick interpretation
if pearson_p < 0.05:
    print("→ Pearson: Significant linear correlation")
if spearman_p < 0.05:
    print("→ Spearman: Significant monotonic correlation")
if kendall_p < 0.05:
    print("→ Kendall: Significant rank correlation")
```

- **Regression**
- Regression: Models relationships between dependent and independent variables.
  o Assumptions: Linearity, no multicollinearity, homoscedasticity.
  • Types of Regression:
  o Linear Regression: Predicts continuous outcomes.
  o Logistic Regression: Predicts binary outcomes.
  o Polynomial Regression: Models nonlinear relationships.
  • Usage:
  o Sales Forecasting: Predict revenue based on past sales.
  o Medical Diagnosis: Predict disease risk based on patient history.

Maximum Likelihood Estimation (MLE)

- **MLE is a technique used to evaluate the importance of samples.**
- It chooses parameter values that best fit the observed data.
- **Process of MLE:**
  - **Initiate a Model:** Select a model that fits the data.
  - **Common Probability Function:** Merges probabilities of all data points using a standard parameter.

- **Maximize Probability:** Find the optimal parameter values where likelihood is maximized.

- **Consistency of MLE:**
  - MLE estimates converge to true parameter values as data increases.
  - Enhances prediction accuracy

Probability in Analytics

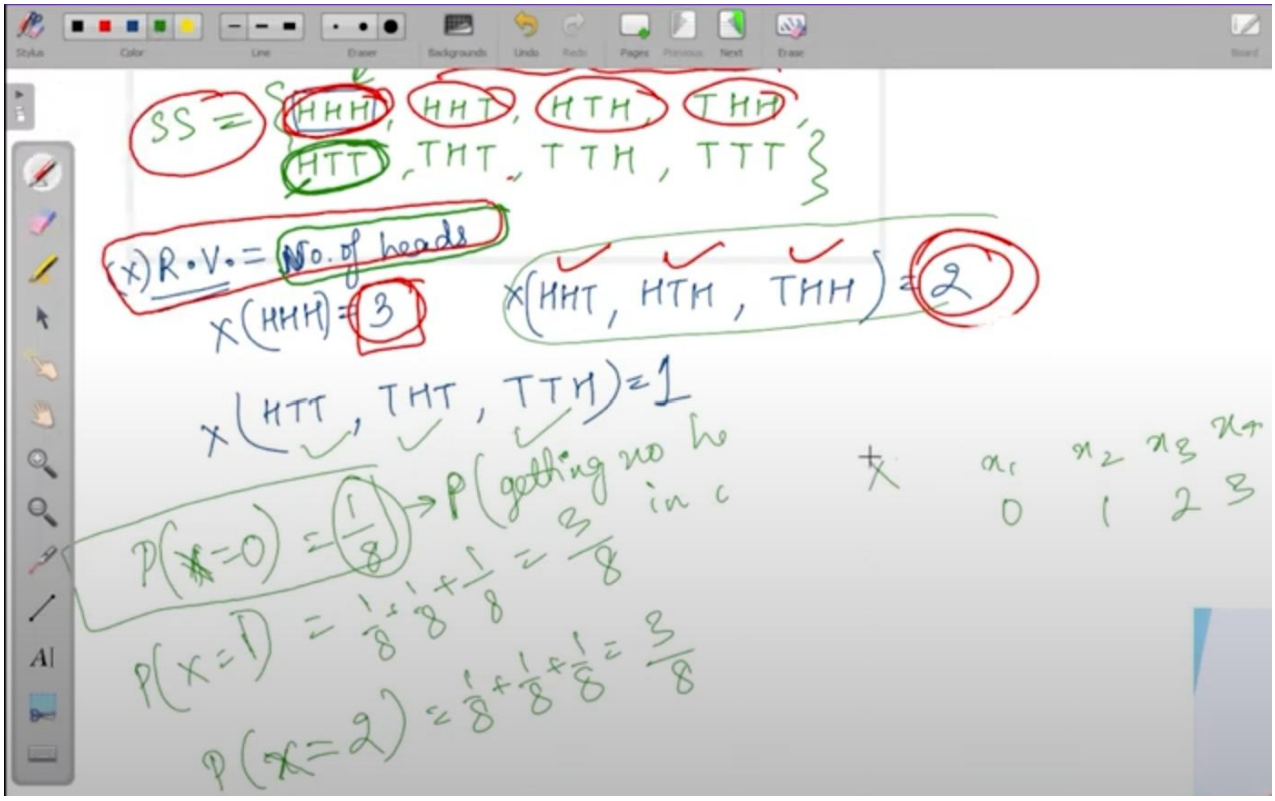
- Used for predicting events, hypothesis testing, and understanding KPI changes.

Essential Concepts

- **Random Experiment:** Experiment where outcomes are uncertain.
- **Sample Space (S):** Set of all possible outcomes.
  - Example: College admission outcome → S = {admitted, not admitted}.
- **Events (E):** Subset of sample space, probability is calculated for these events.
  - Example: Warranty claims less than 10 for 2000 vehicles.

Random Variables

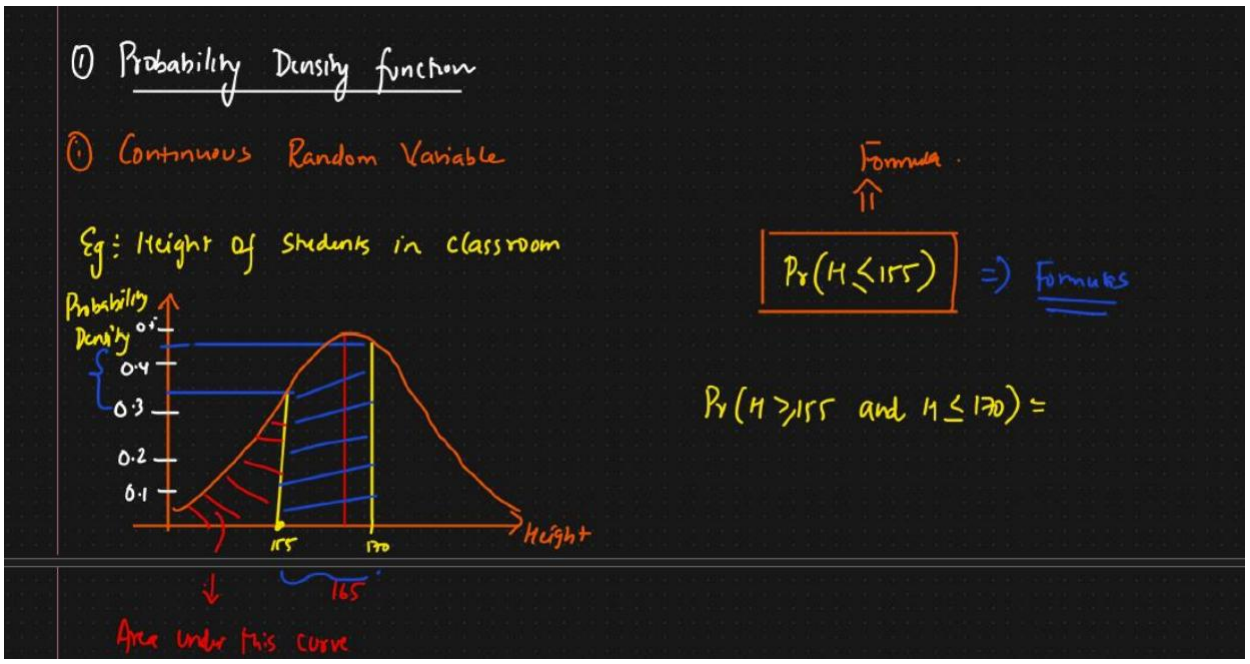
- **Definition:** Function linking sample space outcomes to real numbers.
- **Types:**
  - **Discrete Random Variables:** Finite or countably infinite values.
    - Examples:
      - Credit score
      - Number of e-commerce orders
      - Customer fraud detection
  - **Continuous Random Variables:** Infinite range of values.
    - Examples:
      - Market share variation (0% to 100%)
      - Staff turnover rate
      - Time until system failure



Key Probability Concepts

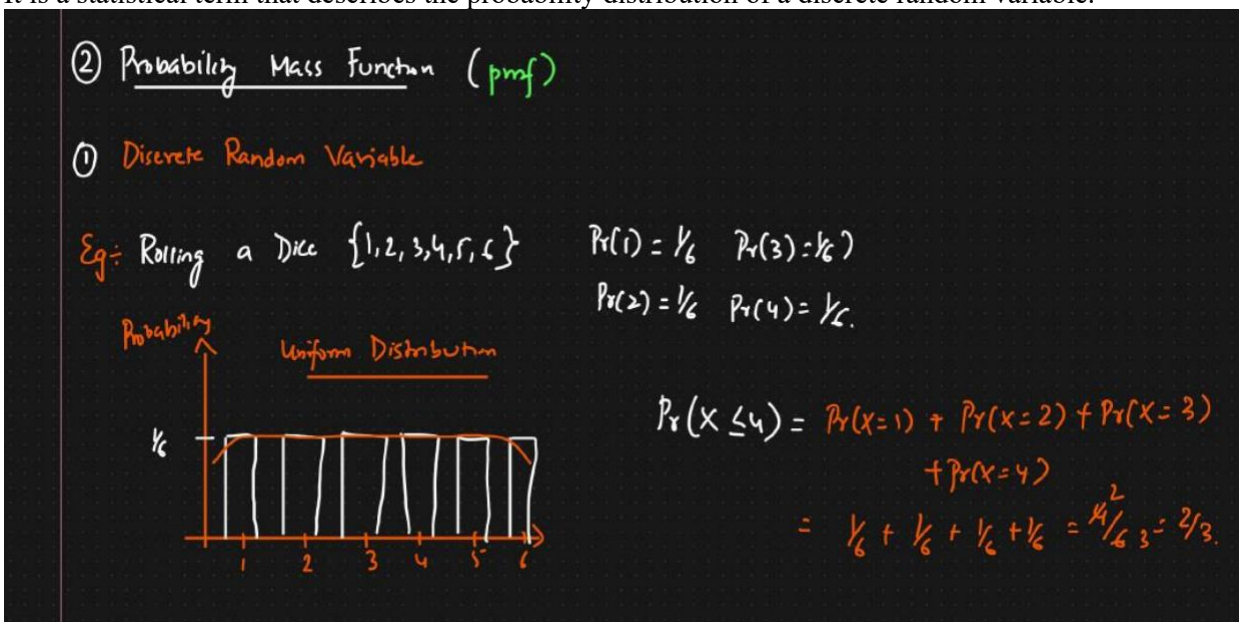
- **Probability Density Function (PDF)**

It is a statistical term that describes the probability distribution of a continuous random variable. The probability associate with a single value is always Zero. Below is the formula for PDF.



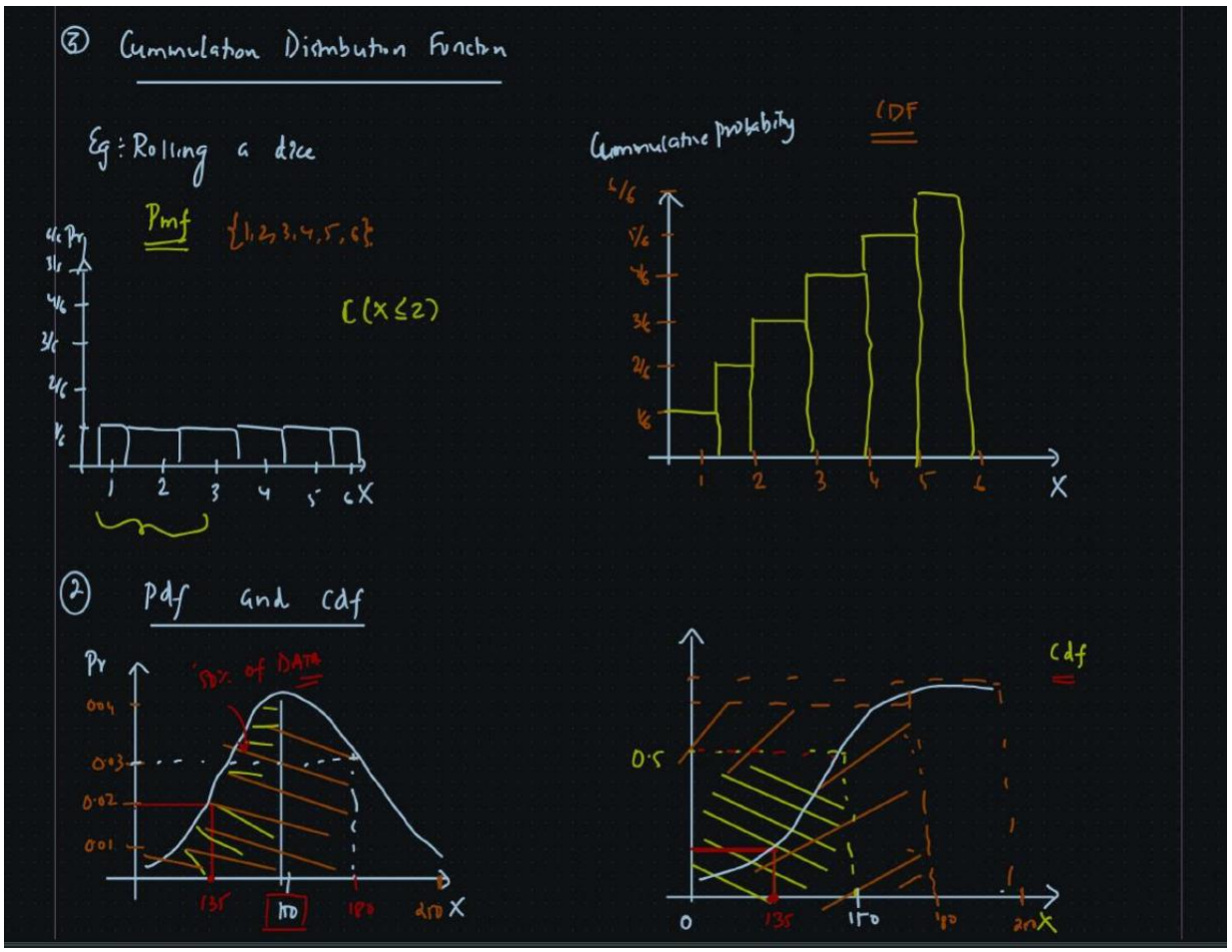
- **Probability Mass Function (PMF)**

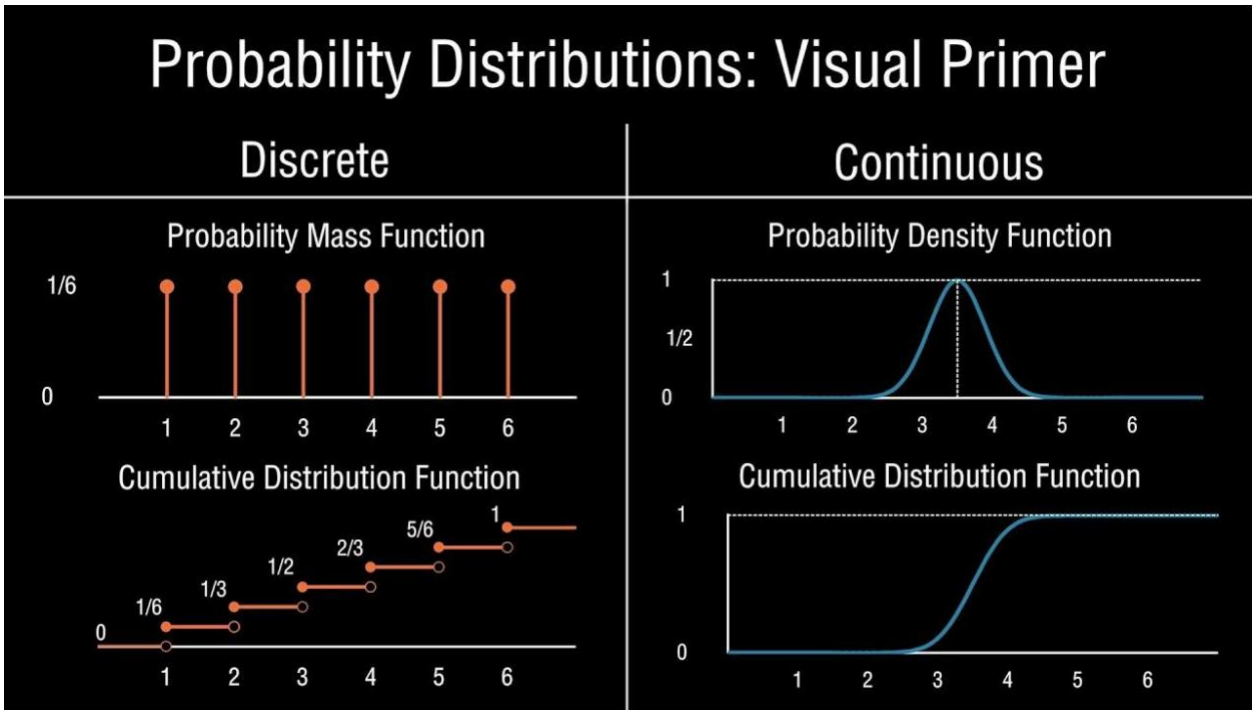
It is a statistical term that describes the probability distribution of a discrete random variable.



- **Cumulative Distribution Function (PDF)**

It is another method to describe the distribution of a random variable (either continuous or discrete).





📘 Probability Distribution

🔹 Definition

A probability distribution describes how the values of a random variable are distributed. It shows the probability of each possible outcome.

🔹 Types of Probability Distributions

1. Discrete Probability Distribution

- Deals with discrete random variables (countable outcomes).

- Sum of all probabilities = 1
- Example: Tossing a coin, rolling a die.

Common Discrete Distributions:

- Binomial Distribution
- Poisson Distribution
- Geometric Distribution

2. Continuous Probability Distribution

- Deals with continuous random variables (uncountable outcomes).
- Represented by a Probability Density Function (PDF).
- Total area under the curve = 1
- Example: Heights of students, time, temperature**.**

Common Continuous Distributions:

- Normal Distribution
- Exponential Distribution
- Uniform Distribution

🔹 **Key Terms**

Term	Meaning
Random Variable	A variable whose value is a result of a random experiment.
PMF	Probability Mass Function (for discrete)
PDF	Probability Density Function (for continuous)
CDF	Cumulative Distribution Function – gives $P(X \leq x)$
Mean (μ)	Expected value or average of the distribution
Variance (σ^2)	Measure of the spread or dispersion of the distribution
Standard Deviation (σ)	Square root of variance

🔹 **Properties of a Probability Distribution**

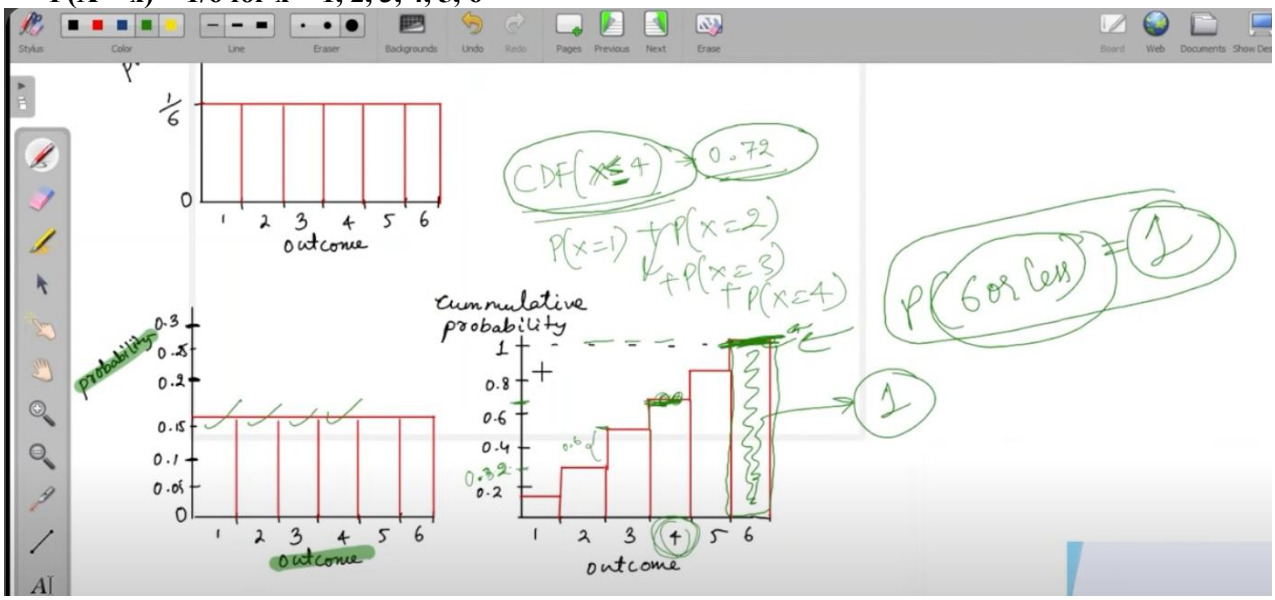
- $0 \leq P(x) \leq 1$ for any value of x
- $\sum P(x) = 1$ (Discrete) or $\int f(x)\,dx = 1$ (Continuous)

🔹 **Examples**

Discrete Example:

Rolling a fair die

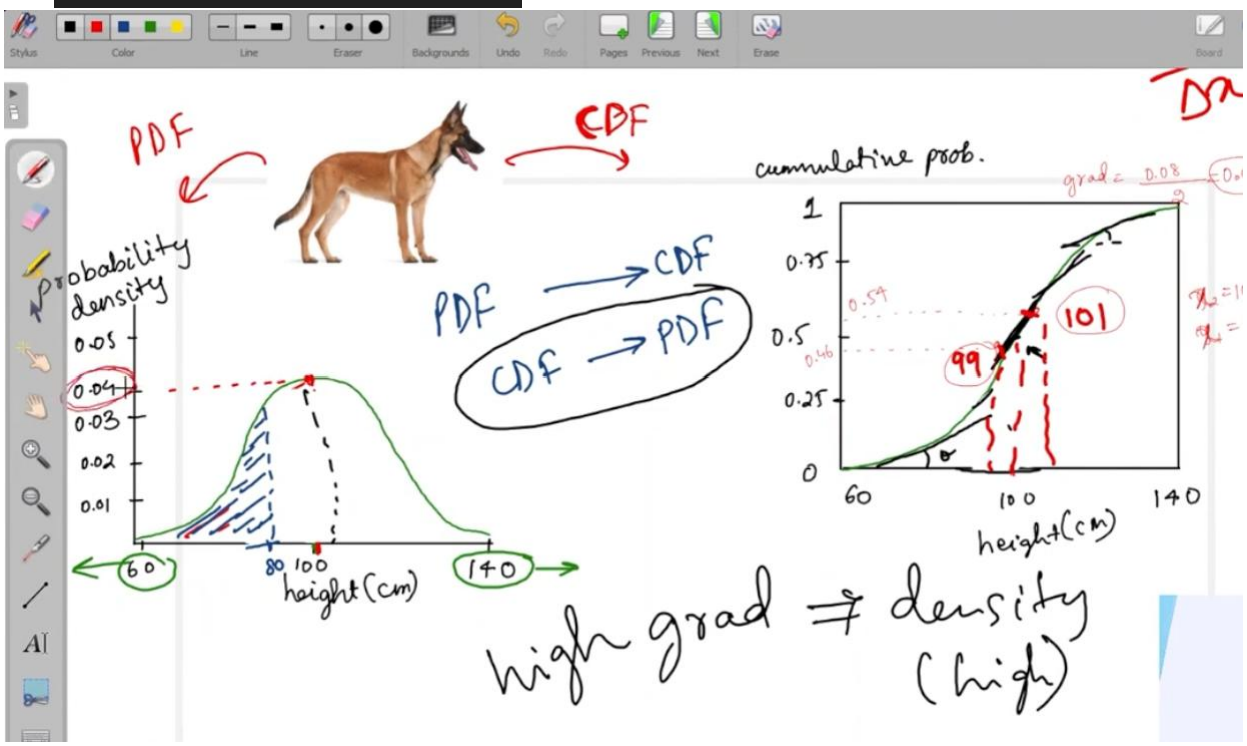
- **Random variable X = number on the die**
- **P(X = x) = 1/6 for x = 1, 2, 3, 4, 5, 6**



Continuous Example:

Let X be the time to complete a task, uniformly distributed between 0 and 10 mins.

- PDF: $f(x) = \frac{1}{10}$ for $0 \leq x \leq 10$
- Mean: $\mu = \frac{a+b}{2} = 5$



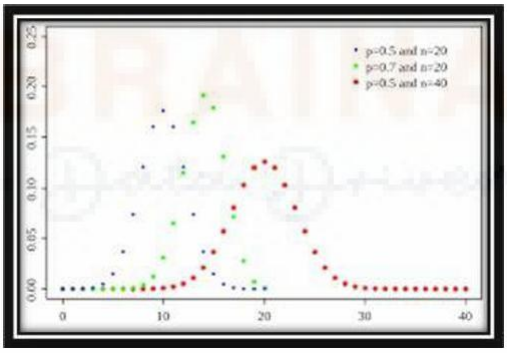
Distribution: Model for Data Classification

- **A distribution framework segments values in a dataset, showing frequency and probability.**
- **Types of Distributions:**
  - **Binomial Distribution**
  - **Multinomial Distribution**
  - **Normal (Gaussian) Distribution**
  - **Uniform Distribution**
  - **Exponential Distribution**
  - **Poisson Distribution**

Types of Probability Distributions

1.Binomial Distribution (Discrete Probability Distribution)

- Used for **repeated binary trials** (e.g., coin flips, product defects).
- Calculates probabilities of success over multiple attempts.
- Probability of success = **p**, failure = **q** (where q = 1 - p).
- Example: Coin flips (Heads or Tails).
- Common in quality control and customer analytics.
- it's concerned with discrete random variables {PMF}
- There are two possible outcomes: true or false, success or failure, yes or no.
- These Experiments is Performs for n trials
- Every trial is an independent trial, which means the outcome of one trial does not affect the outcome of another trial.

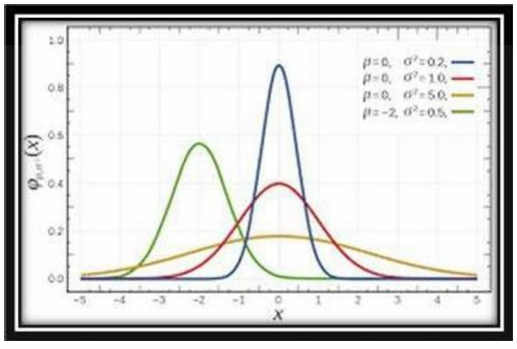


2. Multinomial Distribution

- Used when there are **multiple possible outcomes per trial**.
- Common in **surveys, product classifications, or voting predictions**.

3. Normal (Gaussian) Distribution

- **Bell-shaped curve**, centered around the mean.
- Many natural datasets (e.g., **IQ scores, heights, weights**) follow this distribution.
- it's concerned with Continuous random variables {PDF}
- Normal distributions are symmetrical, but not all symmetrical distributions are normal
  Characteristics of Normal Distribution
  • mean = median = mode
  • Symmetrical about the center
  • Unimodal
  • 50% of values less than the mean and 50% greater than the mean



4. Uniform Distribution

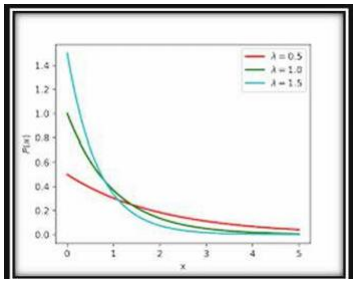
- **All values have equal probability** of occurring.
- Example: Rolling a fair die (each number appears equally).
- **Difference between Normal & Uniform Distribution:**

Aspect	**Uniform Distribution**	**Normal (Gaussian) Distribution**
Shape	Flat, constant probability	Bell-shaped curve
Probability Spread	All values equally probable	Values near the mean are more probable
Examples	Rolling a die, lottery numbers	Heights, IQ scores, weights
Parameters	Min and Max values	Mean and Standard Deviation

---

5. Exponential Distribution

- Models **time between events in a Poisson process** (e.g., machine failure times).
- Probability Density Function (PDF): **λ * e^(-λx), where x ≥ 0**.
- **Applications: Reliability analysis, service time modeling, lifespan prediction.**



6. Poisson Distribution

- Used for modeling **infrequent events occurring over a fixed time period**.
- **Examples:** Customer arrivals, call center requests, accidents.
- Defined by parameter **λ** (average event occurrence).
- it's concerned with discrete random variables {PMF}

Q-Q Plot for Comparing Distributions

- **Quantile-Quantile (Q-Q) Plot:** Visual tool for comparing two distributions.
- If data follows a **Gaussian distribution**, points lie on a straight line.
- Used to check **normality of datasets**.

🔹 **Empirical Rule (68–95–99.7 Rule)**

- 68% of data lies within **1σ** of the mean
- 95% within **2σ**
- 99.7% within **3σ**

🔹 **Properties of Gaussian (Normal) Distribution**

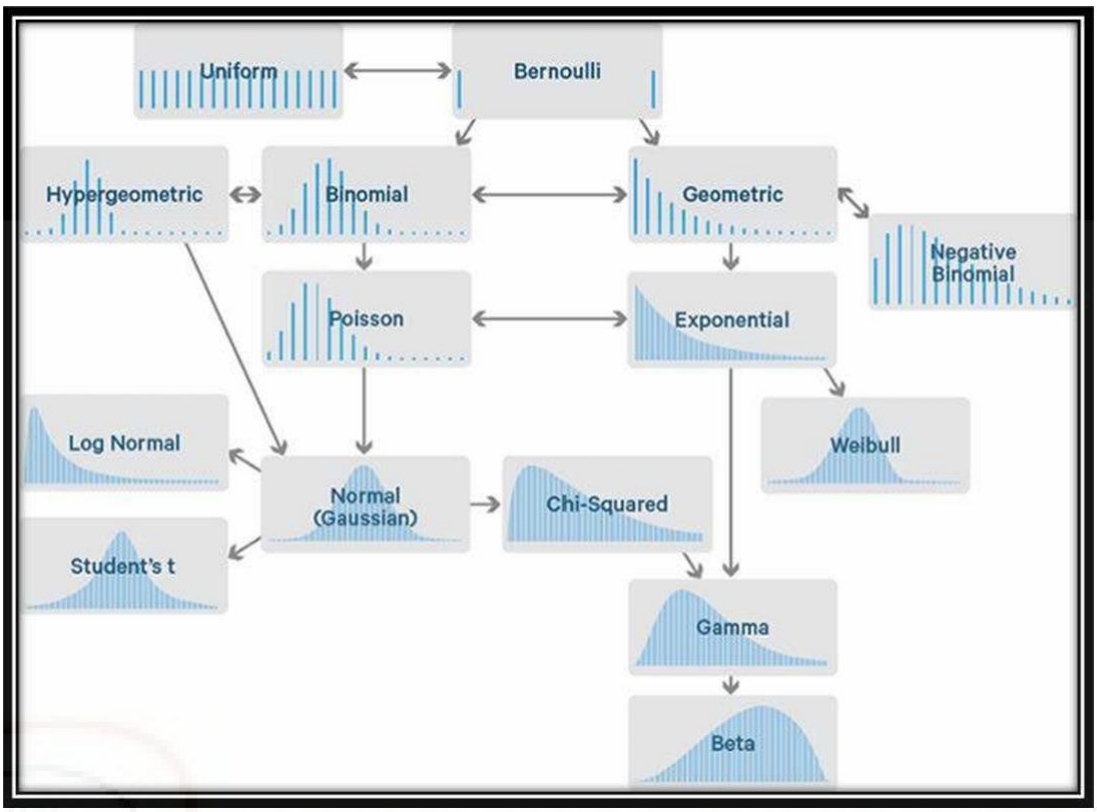
Property	Description
Symmetry	The curve is symmetric about the mean μ\mu
Unimodal	Has a single peak (only one mode)
Mean = Median = Mode	All three central tendency measures are equal
Asymptotic	The tails approach but never touch the x-axis
Bell-Shaped Curve	Characteristic "bell" shape
Total Area = 1	The area under the curve is always 1 (i.e., total probability is 1)
Defined by Mean & Std. Dev.	Entire shape is determined by μ\mu and σ\sigma
Empirical Rule Valid	68-95-99.7 rule applies
Inflection Points	Located at μ±σ\mu \pm \sigma

🔹 **Applications**

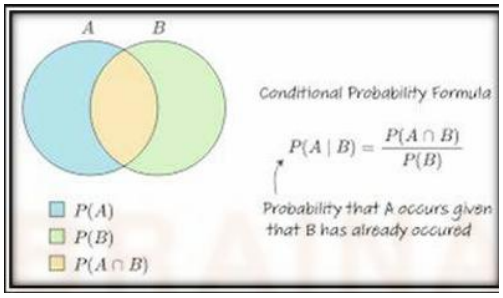
- IQ scores
- Heights and weights
- Stock market returns
- Error measurement in experiments
- Quality control (Six Sigma)

What is the range of values in a standard normal distribution?

- **Theoretical range → From -∞ to +∞.**
- **Most values fall within ±3σ** from the mean (0).
- **Empirical Rule for Standard Normal Distribution (Z-distribution) →**
  - **68%** of values fall within **Z = ±1**.
  - **95%** of values fall within **Z = ±2**.
  - **99.7%** of values fall within **Z = ±3**.
- **Extremely high or low Z-scores are rare**.



Conditional Probability: Probability of an event given prior knowledge of another.



- **Independence & Conditional Independence:**
  - **Independence:** One event does not affect another.
    - Example: Flipping two coins, the first does not affect the second.
  - **Conditional Independence:** Two events are independent given a third event.
    - Example: If we account for weather conditions, two traffic routes may become independent.

1. A/B Testing

1.1 Definition

A/B Testing is a controlled experiment used to compare two versions (A and B) of a product or feature to determine which performs better with respect to a predefined metric.

1.2 Purpose

To enable **data-driven product decisions** by testing changes (UI, pricing, features) on a subset of users before full-scale rollout.

1.3 Key Components

Component	Description
Control Group	The group of users exposed to the existing version (Version A)
Treatment Group	The group of users exposed to the new version (Version B)
Metric	A measurable outcome (e.g., conversion rate, order value, retention rate)

1.4 Steps in A/B Testing

1. **Define the Business Objective**
2. **Formulate Hypotheses**
   - H_0 (Null Hypothesis): No difference between A and B
   - H_1 (Alternative Hypothesis): A significant difference exists
3. **Randomly Assign Users**
   - Typically 50/50 split
4. **Conduct the Experiment**
   - Maintain test duration for statistical significance
5. **Analyze Results**
   - Use appropriate statistical tests
6. **Make a Decision**
   - Accept or reject H_0 based on p-value and business impact

---

1.5 Statistical Tests Used

Data Type	Test
Binary outcome (e.g., conversion)	Z-test for proportions
Continuous metric (e.g., order value)	Independent t-test
Non-normal distribution	Mann-Whitney U test

1.6 Swiggy Example: Button Copy Experiment

Objective: Increase checkout rate
Metric: % of users who place orders

Group	Button Text	Users	Checkout Count	Conversion Rate
A	"Place Order"	10,000	3,200	32%
B	"Get Your Food Now!"	10,000	3,500	35%

- **H_0**: Conversion_A = Conversion_B
- **H_1**: Conversion_B > Conversion_A
- **Test**: Z-test for proportions
- **Outcome**: p-value < 0.05 → Statistically significant difference
- **Decision**: Implement version B

---

2. Sample Size Calculation for A/B Testing

2.1 Why It Matters

Running A/B tests without adequate sample size can lead to false conclusions due to **underpowered** or **overpowered** experiments.

2.2 Formula for Binary Outcomes (Conversion Rate)

$n=2\times(Z1-\alpha/2+Z1-\beta)2\times p(1-p)(p1-p2)2n = \frac{2 \times (Z_{1 - \alpha/2} + Z_{1 - \beta})^2 \times p(1 - p)}{(p_1 - p_2)^2}$

Where:

- nn = sample size per group

- pp = average of p1p_1 and p2p_2
- α\alpha = significance level (usually 0.05)
- β\beta = 1 - power (usually 0.8)
- ZZ = Z-score corresponding to alpha/power

2.3 Example Calculation

- Control conversion: 30%
- Expected uplift: 5% (35%)
- Confidence level: 95%
- Power: 80%

→ Required sample size ≈ **3,000 users per group**

2.4 Tools for Calculation

- Evan Miller's Calculator
- Statsmodels (Python)
- R packages: pwr, power.prop.test

---

3. Common Pitfalls in A/B Testing

3.1 Peeking Too Early

- Stopping the test early when results look promising may cause **false positives**.
- **Best Practice**: Decide the duration/samples **before** starting.

3.2 Unequal User Allocation

- Not randomly assigning users can introduce **selection bias**.
- Use **randomization** or A/B testing platforms to ensure fairness.

3.3 Multiple Testing

- Running too many tests increases **Type I error rate** (false positives).
- Apply **Bonferroni correction** or **false discovery rate control** when necessary.

3.4 Ignoring Practical Significance

- Statistically significant results may not be **business significant**.
- Always evaluate **effect size** and **ROI**.

3.5 External Influences

- Factors like promotions, holidays, or app crashes during the test can skew results.
- Keep external conditions **controlled** or account for them in analysis.

3.6 Poor Metric Choice

- Metrics should be **aligned with business goals** and not easily gamed.
- For example, "add to cart" may not always imply true intent to purchase.

Data Normalization Techniques

Purpose:

To transform data into a standard scale without distorting differences in the ranges of values. It improves the performance of machine learning models and ensures fair treatment of features.

---

1. Min-Max Normalization (Rescaling)

- **Formula:**

x′= x - x_{min}/{x_{max} - x_{min}}

- **Range:** Transforms values to a range [0, 1].
- **Use Case:** Suitable for algorithms like KNN, Neural Networks, or any model that relies on distance.
- **Limitation:** Sensitive to outliers.

---

2. Z-Score Normalization (Standardization)

- **Formula:**

x′={x - \mu}\{sigma}

where μ\mu = mean, σ\sigma = standard deviation.

- **Result:** Mean = 0, Standard Deviation = 1.
- **Use Case:** Effective when data follows a Gaussian distribution; widely used in regression, logistic regression, and clustering.
- **Advantage:** Less affected by outliers compared to Min-Max.

---

3. Robust Scaling

- **Formula:**

x′= {x - \{median}}\{{IQR}}
where IQR = Interquartile Range (Q3 - Q1).

- **Use Case:** When data contains outliers. Maintains robustness by using median and IQR.

---

4. Log Transformation

- **Formula:**

x′=log⁡(x+1)x' = \log(x + 1)

- **Use Case:** Used to handle right-skewed data or reduce the impact of large values.
- **Limitation:** Cannot handle zero or negative values directly.

---

5. Decimal Scaling

- **Formula:**

x′={x}\{10^j}
where jj is the smallest integer such that max(|x'|) < 1.

- **Use Case:** Simple scaling; rarely used in practice today due to more robust methods.

---

6. Unit Vector Scaling (Normalization to Unit Length)

- **Formula:**

x′=x||x||x' = \frac{x} {||x||}

- **Use Case:** Used in text mining (TF-IDF vectors) and cosine similarity where direction is more important than magnitude.

Choosing a Technique:

Situation	**Recommended Method**
Presence of outliers	Robust Scaling, Log Transform
Distance-based models	Min-Max Normalization
Normal distribution assumed	Z-score Standardization
Skewed data	Log or Power Transform
Sparse high-dimensional data	Unit Vector Normalization

S.NO.	Normalization	Standardization
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.
6.	This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
7.	It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
8.	It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

Handling Missing Data When More Than 30% of Values Are Missing

When dealing with a dataset where more than 30% of values are missing, a structured approach is required to ensure data integrity and analysis accuracy. Below are the key steps to handle such scenarios effectively:

1.Understanding the Nature of Missing Data

- **Identify Patterns**: Check if missing values occur randomly or follow a specific pattern.
- **Analyze Missingness Type**:
  - **Missing Completely at Random (MCAR)**: No systematic pattern.
  - **Missing at Random (MAR)**: Missing values depend on observed variables.
  - **Missing Not at Random (MNAR)**: Missing values depend on unobserved variables.

2. Choosing an Appropriate Imputation Method

Basic Imputation Methods

- **Mean/Median Imputation**: Replace missing values with the mean or median of the column.
  - Suitable for numerical data with small percentages of missing values.
  - Not ideal if missing values are high, as it can distort data distribution.
- **Mode Imputation**: Replace missing values with the most frequent value (mode).
  - Suitable for categorical data.

Advanced Imputation Methods

- **K-Nearest Neighbors (KNN) Imputation**:
  - Uses similar data points to estimate missing values based on proximity.
  - Works well when there is a strong correlation between variables.
- **Multiple Imputation**:
  - Generates multiple datasets by imputing different possible values.
  - Reduces uncertainty by incorporating variability.
- **Model-Based Imputation**:

- Uses predictive models (e.g., regression, decision trees) to estimate missing values.
- Suitable for structured datasets with strong relationships between features.

3. Assessing the Impact of Imputation

- Perform **sensitivity analysis** to evaluate how different imputation techniques affect results.
- Compare models with and without imputation to ensure accuracy.

4. Considering Alternatives to Imputation

- **Dropping Rows or Columns**:
  - If a column has more than 50% missing values and is not essential, consider removing it.
  - If a row has excessive missing data and does not contribute significantly, consider removing it.
- **Using Domain Expertise**: Consult experts to make informed decisions about missing values.

Imputation Method	Data Type	Best for	Avoid if
Mean	Numerical (Continuous)	Normally distributed data	Skewed data or outliers
Median	Numerical (Continuous)	Skewed data, outliers present	Symmetric distribution
Mode	Categorical / Ordinal	Categorical data, most frequent value	No dominant category

Why is the median a better measure than the mean?

- The **median** is **less sensitive** to extreme values (**outliers**), making it a more **robust measure** of central tendency.
- **Outliers** can **pull the mean** towards extreme values, **misrepresenting** the dataset.
- The median **remains unaffected** by extreme values, providing a **better representative** measure.
- **Example Dataset**: {1,2,3,4,5,6,7,8,9,100}
  - **Mean** = 15.5 (influenced by the outlier)
  - **Median** = 5.5 (better represents the dataset)
- In **skewed datasets**, the **median** is a **more accurate** measure of central tendency.

What is the meaning of standard deviation?

- **Standard deviation** measures the **spread** or **dispersion** of data points around the mean.
- It indicates how much the data **varies** from the mean.
- It is calculated as the **square root of variance** (average of squared deviations from the mean).
- **Lower standard deviation** → Data points are **closer** to the mean (**less variability**).
- **Higher standard deviation** → Data points are **more spread out** (**greater variability**).
- Expressed in **the same units** as the original data.
- Used to **assess consistency** and **reliability** in statistical analysis.

What is the impact of outliers in a dataset?

Negative Impacts

- **Influence on Central Tendency** → Outliers **pull the mean**, making it **unrepresentative**.

- **Impact on Dispersion Measures** → Increases **standard deviation** and **IQR**, overestimating variability.
- **Skewed Data Distributions** →
  - Positive outliers → **Right-skewed distribution**
  - Negative outliers → **Left-skewed distribution**
- **Misleading Summary Statistics** → Can **distort** statistical interpretation.
- **Impact on Hypothesis Testing** → May lead to **incorrect conclusions**.

Positive Impacts

- **Detection of Anomalies** → Useful in **fraud detection, quality control, and medical research**.
- **Robust Modeling** → Some outliers hold **valuable information**, e.g., **extreme stock price movements** in financial markets.

Methods to screen for outliers in a dataset

- **Box Plots** → Identify outliers **beyond whiskers** in a box-and-whisker plot.
- **Scatterplots** → Detect outliers in **bivariate/multivariate** data.
- **Z-Scores** → Data points with **Z-score > 2 or 3** are potential outliers.
- **IQR (Interquartile Range) Method** →
  - Outliers fall **below Q1 - 1.5 × IQR** or **above Q3 + 1.5 × IQR**.
- **Visual Inspection** → Use **histograms, QQ plots**, etc.

How can you handle outliers in datasets?

- **Data Truncation/Removal** → Remove outliers **only if they are errors**.
- **Data Transformation** → Apply **logarithmic, square root, or inverse transformations** to reduce impact.
- **Winsorization** → Cap extreme values at a **specific percentile** (e.g., **replace top 5% with 95th percentile value**).
- **Imputation** → Replace outliers using **mean, median, or regression-based imputation**.
- **Robust Statistics** → Use **median instead of mean**, **IQR instead of standard deviation**.
- **Model-Based Approaches** → Use models **less sensitive to outliers**, such as **random forests** instead of **linear regression**.
- **Domain Knowledge** → Consult experts before **removing/modifying** outliers.
- **Reporting & Transparency** → Document **how outliers were handled** to ensure **reproducibility**.

Handling Skewed Data

Skewed data can significantly affect the results of statistical analyses, especially those that assume a normal distribution, such as t-tests and linear regression.

1. Data Transformation

Applying mathematical transformations can reduce skewness and help approximate a normal distribution.

Technique	Suitable For	Outcome
Log Transformation	Right-skewed data	Reduces skew
Square Root Transformation	Moderate skew	Reduces skew
Box-Cox Transformation	Positive values	Flexible normalization
Yeo-Johnson Transformation	Positive or negative values	Normalization

Example: Applying log(order_value + 1) can reduce the effect of extreme values in food delivery price data.

---

2. Remove or Cap Outliers

Outliers can heavily influence skewness. Two common methods to address them are:

- Winsorization: Caps extreme values at a defined percentile (e.g., 95th percentile).
- Trimming: Removes a specified percentage of the lowest and highest values.

Example: Capping unusually high delivery times at the 99th percentile before analysis.

---

3. Use Robust Statistical Methods

When the data is skewed, it is recommended to use statistical measures and tests that are less sensitive to outliers.

Common Measure	Robust Alternative
Mean	Median
Standard Deviation	Interquartile Range (IQR)
T-test	Mann-Whitney U Test (independent groups), Wilcoxon Signed-Rank Test (paired groups)

---

4. Modeling Techniques for Skewed Data

Certain models are better suited for skewed data:

- Tree-based models (e.g., Random Forest, XGBoost) do not assume normality.
- Robust regression techniques can be used to reduce the impact of outliers.

---

5. Visual and Quantitative Assessment

- Use histograms and boxplots to visually inspect skewness.
- Calculate the skewness coefficient:
  - Greater than +1: Highly right-skewed
  - Less than -1: Highly left-skewed
  - Between -1 and +1: Approximately symmetric

Example in Practice (Swiggy Scenario)

Objective: Analyze delivery time across two regions.

- Raw delivery time data shows right skew due to a few extreme delays.
- Apply log transformation: log(delivery_time + 1) to reduce skew.
- Use median rather than mean to report central tendency.
- Use Mann-Whitney U test instead of t-test to compare two groups.

Aspect	Univariate Analysis	Bivariate Analysis	Multivariate Analysis
Nature	Examines a single variable.	Examines the relationship between two variables.	Analyzes multiple variables simultaneously.
Focus	Analyzing distributions, summary statistics, and characteristics.	Focuses on how changes in one variable are associated with changes in another variable.	Observes how multiple variables interact and influence each other.
Examples	Histograms, Box plots, Mean, Median, Standard deviation.	Scatter plots, Correlation coefficients, cross-tabulations.	Pairplot, Principal Component Analysis (PCA), Factor Analysis.
Application	Useful for understanding the characteristics of a single variable.	Useful for exploring relationships between two variables.	Useful for a comprehensive understanding of interactions among multiple variables.

What is the meaning of KPI in statistics?

Ans. KPI stands for Key Performance Indicator.

KPIs are simple indicators or metrics used to measure and evaluate the performance of a project, system or company.

Definition: KPIs are specific indicators designed to measure effectiveness and efficiency in various factors in a company or brand.

Applications:-KPIs in business, finance, healthcare, education, etc. It provides software packages in var ious fields.

The selection of KPI is often based on the company's specific goals and measured indicators or standards.

Monitoring and Analysis: Regular monitoring and analysis of KPI can provide valuable information by helping businesses identify areas for improvements identified by statistics and measures progress against goals!!!
KPI examples:-In business: revenue growth rate, customer acquisition rate (CAC), customer retention rate.-Health care: patient satisfaction; life expectancy// readmission rates...
-In education: overall student performance, graduation rates, teacher training.

How do you calculate the sample size needed?
Ans. To decide the specified pattern size for you examine:
Define Research Objectives: Clearly outline your studies dreams and questions.
Set Significance Level and Margin of Error: Choose an important stage (α) and determine the ideal margin of mistakes (E).
Estimate Population Variability: Estimate the population variability (σ) or use conservative estimates if genuine values are unavailable.
Determine Population Size: Identify the total population size (N) below consideration.
Select Sampling Type: Choose between random or stratified sampling, depending to your examine layout.
Choose Statistical Test: Select the correct statistical test or analysis in your studies.
Apply Sample Size Formula or Software: Utilize a sample length method or devoted software program gear to calculate the specified pattern size.
Consider Practical Constraints: Account for practical constraints and capacity non-reaction with the aid of adjusting the calculated pattern size.
Conduct Study and Analyze Data: Execute the examination, collect records from the determined sample length, and perform the chosen analysis.
Interpret Results: Analyze outcomes and draw significant conclusions based totally on the achieved pattern size.
Sample length calculations are essential to make certain your examine generates sufficient records for significant conclusions at the same time as retaining manipulate over mistakes and precision