Chapter 2: Statistical Foundation (8 hrs)

1. Statistics:

Statistics is the branch of mathematics that deals with collecting, organizing, analyzing, interpreting, and presenting data**.** It helps us understand large volumes of data and draw meaningful conclusions. In simple terms, Statistics is the science of learning from data.

Suppose you are analyzing the performance of a computer network. You collect data on packet delays, loss rates, and throughput. Using statistical tools like *mean, variance, and correlation*, you can identify:

- Whether packet delay is increasing,
- Which factor affects throughput most,
- And whether the changes are statistically significant.

Key Functions of Statistics

1. **Data Collection:** Gathering relevant data from surveys, experiments, or observations.
2. **Data Organization:** Arranging data using tables, charts, or graphs.
3. **Data Analysis:** Applying mathematical methods (mean, correlation, regression, etc.) to summarize or explore relationships.
4. **Interpretation:** Drawing conclusions and making informed decisions based on data patterns.

Why Statistics is Important in Data Analysis

Data analysis is the process of **extracting useful information** from raw data. Statistics provides the **theoretical foundation and tools** for this process.

Importance:

1. **Understanding Data Patterns**
    - Helps identify trends, variations, and anomalies in data.

- Example: Average student marks or monthly sales trends.

2. **Summarizing Large Data**
    - Statistical measures like *mean, median, mode, standard deviation* summarize complex data in simple terms.
3. **Making Decisions Under Uncertainty**
    - In real life, we often make decisions with incomplete data. Statistics (through probability, confidence intervals, hypothesis testing) helps us estimate and make reliable predictions.
4. **Evaluating Relationships**
    - Techniques like *correlation* and *regression* show how variables are related e.g., "Does study time affect exam scores?"
5. **Testing Hypotheses**
    - Helps decide whether an observed pattern is statistically significant or occurred by **chance**.
6. **Foundation for Machine Learning & Data Mining**
    - Most data-driven technologies (AI, data science, business analytics) rely on statistical models to train algorithms and make predictions.

1. Types of Variables: Numeric and Categorical

Variable:

A **variable** is a **characteristic or attribute** of an object, event, or individual that **can take different values**. Variables represent **properties of objects, events, or people** that can change or vary from one observation to another.

- Variables are central to **data analysis** because they define **what we are measuring, observing, or predicting**.
- Every dataset contains **rows (observations)** and **columns (variables/features)**.

Example:

- In a class: student height, weight, age, gender, marks in a test.
- In a network: packet delay, throughput, error rate.

Types of Variables

Variables are mainly classified into **two broad categories**: **Numeric (Quantitative)** and **Categorical (Qualitative)** types.

a) Numeric (Quantitative) Variables

- Represent measurable quantities expressed as **numbers**.
- We can perform **mathematical operations** (addition, subtraction, averaging).
- Often visualized using **histograms, box plots, scatter plots**.

Subtypes:

1. ***Continuous Variables***
    - Can take **any value within a range**, including decimals.
    - **Used in:** scientific measurement, performance metrics.
    - **Visualization:** Histograms, line plots
    - **Example:** Height = 165.3 cm, Temperature = 37.2°C
2. ***Discrete Variables***

- Can take only **specific countable values**.
- **Used in:** counting events or objects.
- **Visualization:** Bar charts, frequency tables
- **Example:** Number of students in a class = 40, Number of errors = 3

b) Categorical (Qualitative) Variables

- Represent **categories or labels** rather than numbers.
- You **cannot perform arithmetic operations** directly on them.
- Often visualized using **bar plots, pie charts, or frequency tables**.

Subtypes:

1. ***Nominal Variables***
   - Categories **without any natural order**.
   - Useful for **classification** or grouping.
   - **Visualization:** Bar charts, Pie charts, etc
   - **Examples:**
     - Gender → Male, Female
     - Blood Type → A, B, AB, O
     - Department → CSE, IT, CE
2. ***Ordinal Variables***
   - Categories **with a meaningful order or ranking**, but differences between ranks may not be uniform.
   - **Visualization:** Bar charts or ordered plots
   - **Examples:**
     - Exam grades → Poor, Average, Good, Excellent
     - Customer satisfaction → Very Unsatisfied → Very Satisfied
     - Skill levels → Beginner, Intermediate, Expert

2. Empirical Distribution

- In data science, the word "empirical" means "observed". Empirical distributions are distributions of observed data, such as data in random samples.
- An **empirical distribution** represents how data are **actually distributed** based on **observed (real-world) data**, rather than on a theoretical model (like the normal or binomial distribution).
- In simple words, **Empirical distribution shows what the data tell us**, not what we assume.
- An **empirical distribution** is the **actual distribution** of observed data. It tells us **how often** each value (or range of values) occurs in a dataset — based purely on **real observations**, not theory.
- It is called *empirical* because it comes from **experience or experiment** (the word "empirical" means "based on observation").
- Analogy: Imagine you are a teacher checking the **exam scores** of 10 students. You don't assume what the scores *should be* (like "scores are normally distributed") — you just **look at what they are**. The pattern of those actual scores is your **empirical distribution**.
- Significance:
    - It shows **the frequency or proportion of different values** in your dataset.
    - Helps understand the **pattern, variability, and trends** in data.

Definition:

The **empirical distribution** of a dataset is the **distribution of observed values**. It shows how frequently each value (or range of values) occurs in the data.

Instead of assuming a theoretical model (like Normal, Exponential, Binomial), the empirical distribution **comes directly from the sample itself**.

Mathematically, if you have a dataset of n observations: x1, x2, x3, ..., xn

Then the **empirical distribution function (EDF)** is defined as:

$$F(x) = \frac{\text{Number of observations } \leq x}{n}$$

This is called the **Empirical Distribution Function (EDF)**.

This distribution **does not assume any parametric form**. It simply represents "what the data actually looks like".

Simple Example:

Suppose you have the ages of 6 people:

Data: [12, 15, 16, 16, 20, 30]

For x = 16:

Count values ≤ 16:

→ (12, 15, 16, 16) = **4** values

$$\hat{F}(16) = \frac{4}{6} = 0.67$$

Interpretation: **67% of people in this sample are aged 16 or below.**

Numerical Example:

Suppose the marks (out of 10) obtained by 10 students are:

Student	Marks
1	4
2	6
3	7
4	5
5	6
6	8
7	6
8	9
9	7
10	5

Step 1: Create a Frequency Table

Marks	Frequency	Relative Frequency
4	1	1/10 = 0.10
5	2	2/10 = 0.20
6	3	3/10 = 0.30
7	2	2/10 = 0.20
8	1	1/10 = 0.10
9	1	1/10 = 0.10

The **relative frequency** column shows the **empirical probability** of each mark. For example:

- Probability that a student scores 6 = 0.30
- Probability that a student scores ≤ 6 = 0.10 + 0.20 + 0.30 = 0.60

This is the **Empirical Distribution Function (EDF)**:

$$F(x) = \frac{\text{Number of observations } \leq x}{n}$$

Step 2: Visualize with a Histogram

If you draw a **histogram** with marks on the x-axis and frequency on the y-axis, the shape you see — that's the **empirical distribution** of marks.

Interpretation

- The most frequent mark = **6**
- About **60%** of students scored **6 or below**.
- The data shows the **actual pattern of student performance**, not a theoretical one.

Numerical:

The table below shows the number of hours that 10 students studied for an exam.

Student	Hours Studied
1	2
2	3
3	3
4	4
5	4
6	4
7	5
8	6
9	6
10	8

a) Construct the **empirical frequency distribution** of the data.

b) Compute the **empirical cumulative distribution function (EDF)**.

c) Interpret the distribution briefly.

Solution:

(a) Frequency and Relative Frequency Table

Hours (x)	Frequency (f)	Relative Frequency (f/n)
2	1	1/10 = 0.10
3	2	2/10 = 0.20
4	3	3/10 = 0.30
5	1	1/10 = 0.10
6	2	2/10 = 0.20
8	1	1/10 = 0.10

Total n =10

(b) Empirical Cumulative Distribution Function (EDF)

Hours (x)	Cumulative Frequency (≤x)	F(x) = (CF/n)
2	1	0.10
3	3	0.30
4	6	0.60
5	7	0.70
6	9	0.90
8	10	1.00

(c) Interpretation

- The **most common study duration** (mode) is **4 hours**.
- **60% of students** studied **4 hours or less**.
- **10% of students** studied **8 hours**, the highest value.
- The empirical distribution shows that **most students studied between 3 to 5 hours**

a) Numeric data: histograms, normal, exponential, power laws

i) Histograms:

A histogram is a type of graphical representation used in statistics to show the distribution of numerical data.

A histogram is a **bar graph** that shows **how many data points fall into different ranges (bins)**.

It looks somewhat like a bar chart, but unlike bar graphs, which are used for categorical data, histograms are designed for continuous data, grouping it into logical ranges, which are also known as "bins."

It divides the data into **intervals** (called **bins**) and shows how many data points fall into each bin. In a histogram, data is grouped into continuous number ranges, and each range corresponds to a vertical bar.

Key Features

- Used for **continuous** (or very large discrete) numeric data.

- X-axis → **Bins (ranges of values)**.
- Y-axis → **Frequency** (count of data points in each bin).
- Bars **touch each other**, showing continuity.

Parts of a Histogram

1. **Bins** – continuous intervals (e.g., 0–5, 5–10)
2. **Frequencies** – counts
3. **Bars** – height = frequency
4. **Range** – from minimum to maximum values
5. **Class width** – size of each bin

Example 1:

Heights of **10 students** (in cm): 144, 148, 150, 155, 158, 160, 162, 170, 172, 175

Let bins be:

Bin (cm)	Frequency
140–150	2
150–160	4
160–170	2
170–180	2

Histogram bars are drawn with heights 2, 4, 2, 2.

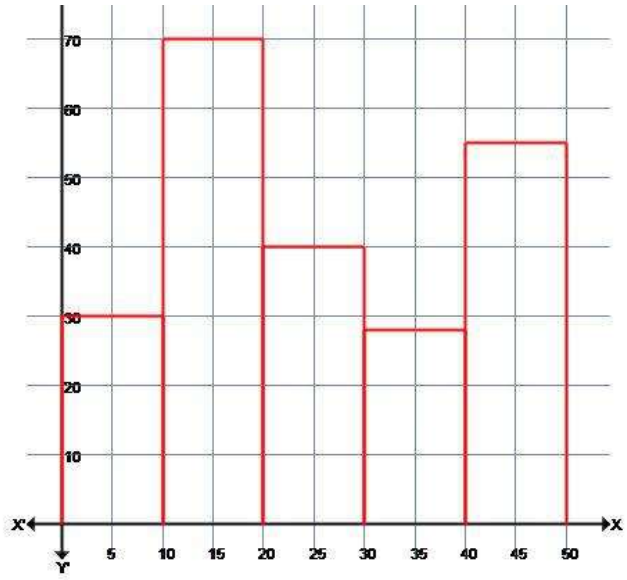
Example 2:

Present the following information as a histogram:

Marks	0-10	10-20	20-30	30-40	40-50
No. of students	30	70	40	28	55

Solution:

We take the Marks on the graph's horizontal axis and, based on the first column of the data, set the scale to 1 unit = 10. We pick number of students on the vertical axis of *the graph and use the second column of the table to* determine the scale: 1 unit = 10. Now we'll create the relevant histogram.



Types of Histogram

There are various variations of the histograms based on their shapes:

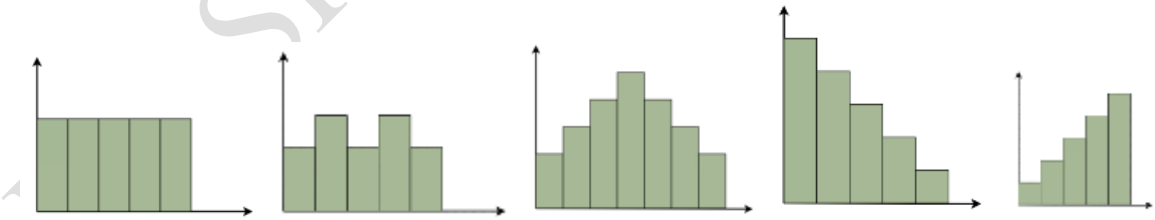


Figure: a) Uniform b)Bimodal c)symmetric d) Left skewed e)right skewed

a) Uniform Histogram

A Uniform Histogram shows uniform distribution means that the data is uniformly distributed among the classes, with each having a same number of elements. It may display many peaks, suggesting varying degrees of incidence.

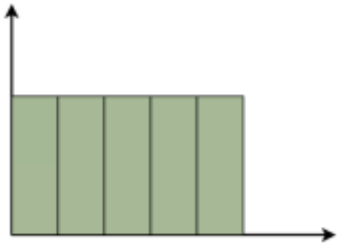


Figure: Uniform Histogram

b) Bimodal Histogram

A histogram is called bimodal if it has two distinct peaks. This implies that the data consists of observations from two distinct groups or categories, with notable variations between them.

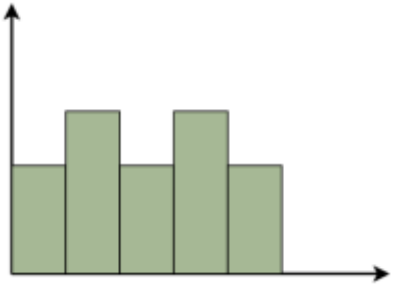


Figure: Bimodal Histogram

c) Symmetric Histogram

Symmetric Histogram is also known as a bell-shaped histogram, it has perfect symmetry when divided vertically down the centre, with both sides matching each other in size and shape. The balance reflects a steady distribution pattern.

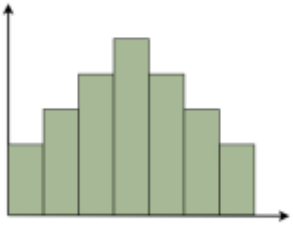


Figure: Symmetric Histogram

d) Right-Skewed Histogram

A right-skewed histogram shows bars leaning towards the right side.

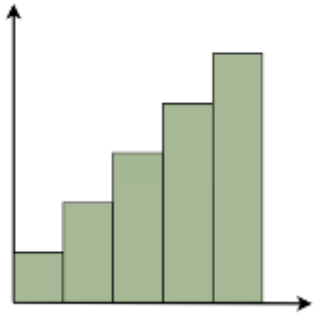


Figure: Right-Skewed Histogram

e) Left-Skewed Histogram

A left-skewed histogram shows bars that lean towards the left side.

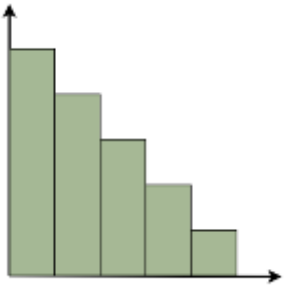


Figure: Left-Skewed Histogram

Frequency Histogram

A frequency histogram visually displays, frequency distribution of how often specific values appear in data.

Example: Let's say we have the ages of 12 people: Ages: [12, 15, 17, 18, 18, 19, 21, 22, 24, 25, 25, 26]

Frequency Table

Range(Bins)	Frequency
10-14	1
15-19	4
20-24	3

Range(Bins)	Frequency
25-29	4

Relative Frequency Histogram

Relative Frequency Histogram displays proportions instead of exact counts for each interval.

Example: Given the test scores of 10 students: Scores: 55, 60, 62, 70, 75, 78, 80, 82, 85, 90

Relative Frequency Table

Interval (Bins)	Frequency	Relative Frequency
50–59	1	1/10 = 0.10
60–69	2	2/10 = 0.20
70-79	3	3/10 = 0.30
80-89	3	3/10 = 0.30
90-99	1	1/10 = 0.10

Cumulative Frequency Histogram

A cumulative frequency histogram is a graph that depicts the total number of values up to a specific point.

The cumulative frequency table below shows the distribution of test scores for 10 students:

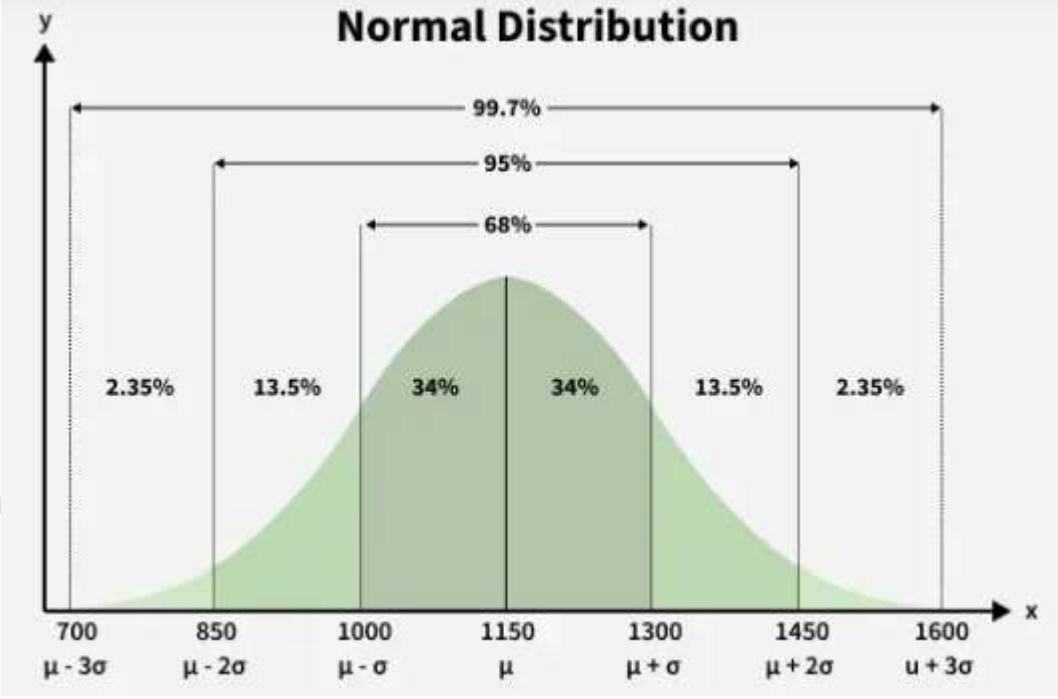
Interval	Frequency	Cumulative Frequency
50-59	1	1
60-69	2	1 + 2 = 3

Interval	Frequency	Cumulative Frequency
70-79	3	3 + 3 = 6
80-89	3	6 + 3 = 9
90-99	1	9 + 1 = 10

ii) Normal Distribution:

Normal Distribution is the most common or normal form of distribution of Random Variables, hence the name "normal distribution." It is also called the Gaussian Distribution in Statistics or Probability.

A symmetric distribution where most observations are near the **mean**, and very few are extremely high or low.



Properties

- Mean ≈ Median ≈ Mode
- Symmetric
- 68% data lies within ±1 standard deviation

- Many natural phenomena: height, blood pressure, errors

Numeric Example

Suppose exam marks of 20 students cluster around 60:

Marks: 45, 50, 52, 55, 58, 58, 60, 60, 61, 62, 62, 63, 65, 65, 66, 68, 70, 72, 75, 80

The histogram will show:

- Low at extremes (45, 80)
- Highest around 58–65

That is a near-normal distribution.

Draw the figure yourself.

iii) Exponential distribution

The Exponential Distribution is one of the most commonly used probability distributions in statistics and data science. It is widely used to model the time or space between events in a Poisson process.

The **exponential distribution** describes the **time between events** that happen **continuously and independently** at a constant average rate.

In simple terms, it describes how long you have to wait before something happens, like a bus arriving or a customer calling a help center.

The exponential distribution depends on one parameter, called λ (lambda) — the rate of events.

$$f(x;\lambda) = \lambda e^{-\lambda x},\ x \geq 0$$

Where:

- λ= **average number of events per unit time (the *rate*)**
- x= **time between events**
- e= **Euler's number (≈ 2.718)**
- $f(x;\lambda)$ =**probability density function (PDF)**

- **Mean (μ)=** $\frac{1}{\lambda}$
- **Standard *Deviation (σ)*** $= \frac{1}{\lambda}$

Imagine you are **waiting for a bus** that arrives on average **every 10 minutes**, but randomly. You often wait:

- 1–5 minutes: very common
- 10–20 minutes: less common
- 30+ minutes: very rare

This matches an exponential distribution.

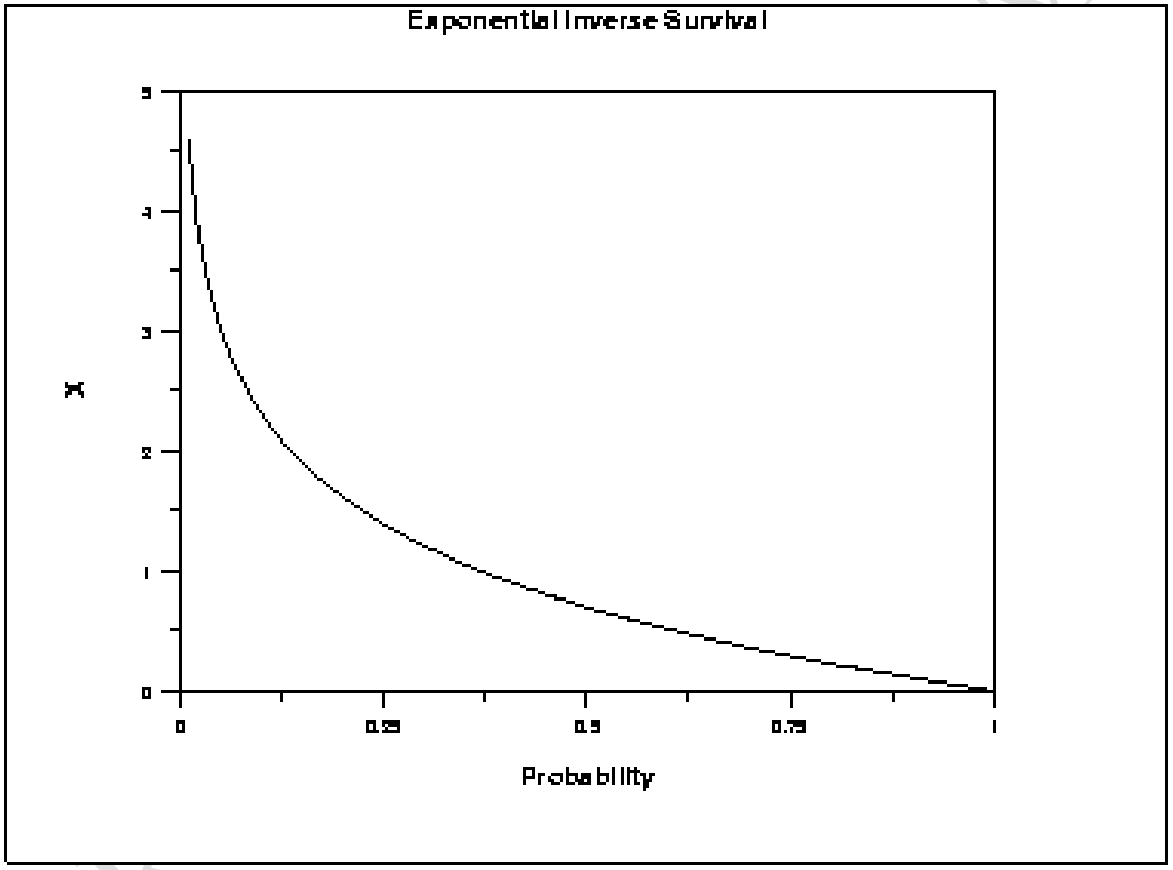


Figure: Exponential distribution

Cumulative Distribution Function (CDF)

- The CDF gives the probability that the waiting time is less than or equal to some value x:

$$F(x;\lambda) = 1 - e^{-\lambda x}$$

- If λ is large, things happen fast (short waits).
- If λ is small, things happen slowly (long waits).
- The curve shows that short waits are common, but long waits occasionally happen

Numerical:

1. **A machine fails on average** $\lambda = 1$ time per 5 hours**. What is the probability the machine fails** within 3 hours**?**

   Solution:

   First find λ in events/hour:

$$\lambda = \frac{1}{5} = 0.2$$

   The probability the machine fails **within 3 hours:**

$$P(X \leq t) = 1 - e^{-\lambda t}$$

$$= 1 - e^{-0.2 \times 3}$$

$$= 1 - e^{-0.6}$$

$$e^{-0.6} \approx 0.548$$

$$P = 1 - 0.548 = 0.452$$

Answer: There is a 45.2% chance the machine fails within 3 hours.

2. **In a computer server system, requests arrive randomly at an average rate of 30 requests per minute. Assume the time between incoming requests follows an Exponential Distribution.**

Calculate:

1. **The probability that the next request arrives within 2 seconds.**
2. **The probability that the next request arrives after 5 seconds.**
3. **The probability that the time between two requests lies between 2 and 4 seconds.**

Solution:

Convert rate to per second

Given:
30 requests per minute

$$\lambda = \frac{30}{60} = 0.5 \text{ requests per second}$$

So $\lambda = \mathbf{0.5}$.

Probability request arrives within 2 seconds

Formula (CDF):

$$P(T < t) = 1 - e^{-\lambda t}$$

$$P(T < 2) = 1 - e^{-0.5 \cdot 2} = 1 - e^{-1}$$

$$= 1 - 0.3679 = 0.6321$$

Probability = 0.6321 (≈ 63.21%)

Probability request arrives after 5 seconds

Use survival function:

$$P(T > t) = e^{-\lambda t}$$

$$P(T > 5) = e^{-0.5 \cdot 5} = e^{-2.5}$$

$$\approx 0.0821$$

Probability = 0.0821 (≈ 8.21%)

Probability time is between 2 and 4 seconds

$$P(2 < T < 4) = P(T < 4) - P(T < 2)$$

Compute each:

(i) $P(T < 4)$

$$P(T < 4) = 1 - e^{-0.5 \cdot 4}$$

$$= 1 - e^{-2}$$

$$= 1 - 0.1353 = 0.8647$$

(ii) $P(T < 2)$

From earlier:

$$P(T < 2) = 0.6321$$

Combine:

$$P(2 < T < 4) = 0.8647 - 0.6321 = 0.2326$$

Probability = 0.2326 (≈ 23.26%)

Final Answers

Part	Result
(1) Probability next request arrives within 2 seconds	**0.6321**
(2) Probability next request arrives after 5 seconds	**0.0821**
(3) Probability time lies between 2 and 4 seconds	**0.2326**

Assignment:

Suppose that the longevity of a light bulb is exponential with a mean lifetime of eight years. If a bulb has already lasted 12 years, find the probability that it will last a total of over 19 years.

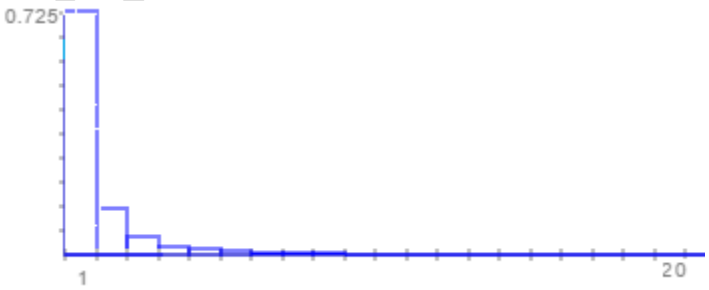
iv) Power-law distribution

The **power law** (also called the scaling law) states that a relative change in one quantity results in a proportional relative change in another. The simplest example of the law in action is a square; if you double the length of a side (say, from 2 to 4 inches) then the area will quadruple (from 4 to 16 inches squared). A power law distribution has the form $Y = k X^{\alpha}$, where:

- X and Y are variables of interest,
- α is the law's exponent,
- k is a constant.

Any inverse relationship like $Y = X^{-1}$ is also a power law, because a change in one quantity results in a negative change in another.

- The power law describes systems where a few things are very big, and many things are very small — but those big things still play a *massive role*.
- In simple words: Small events are common, big events are rare.
- It is extremely important in data science, economics, and network analysis because it helps us understand systems with a few very large values and many small ones (known as *heavy-tailed* or *long-tail* distributions).
- In normal linear plots, power laws appear *curved* with a long tail.



- A cluster of values dominates at one end of the graph.

A special type of this type of distribution is the Pareto Principle (also called the Pareto Law), which is an unscientific "law" that states *80% of effects come from 20% of causes*. In other words, most of what we do has little effect.

Real-Life Examples

- **City populations:** Many small towns, few huge cities
- **Wealth distribution:** Many people have small wealth, few are billionaires
- **Word frequency in texts:** "the", "is" appear often, long rare words appear rarely
- **Internet:** Few websites have millions of visitors, most have very few
- **Earthquakes:** Many small tremors, few big earthquakes

Numerical Example:

Suppose we have **city populations** (in thousands) and we approximate a power-law with **$\alpha = 2$**:

City Size (x)	Probability ($P(x) \propto 1/x^2$)
1	$1/1^2 = 1.0$
2	$1/2^2 = 0.25$
3	$1/3^2 \approx 0.111$
4	$1/4^2 = 0.0625$
5	$1/5^2 = 0.04$

Interpretation:

- Small cities (1k population) are very **common**
- Large cities (5k population) are **rare**

b) Categorical data: bar plots, binomial distribution, Zipf's law

What is Categorical Data?

Categorical data is data that can be **grouped into categories**.

- It **cannot be measured on a scale** like numbers.
- It is usually **counted** instead of calculated.
- Each data point belongs to **one category**.

Examples:

- Colors: Red, Blue, Green
- Fruits: Apple, Banana, Mango
- Gender: Male, Female, Other
- Yes/No responses
- Car brands: Toyota, BMW, Honda

2. Types of Categorical Data

a) Nominal Data

- Categories have **no specific order**
- Example: Eye color (Blue, Brown, Green), Types of fruits

b) Ordinal Data

- Categories have a **specific order**, but differences are **not measurable**
- Example: Movie ratings (Poor, Average, Good, Excellent), Education level (High School, Bachelor, Master)

How to Represent Categorical Data?

a) Frequency Table

- Shows how many data points fall in each category

  Example: Favorite fruit survey (10 students)

Fruit	Count
Apple	3
Mango	4

Fruit	Count
Banana	2
Orange	1

b) Bar Plot

- Visual representation of category frequencies
- Height of bar = number of occurrences

c) Pie Chart

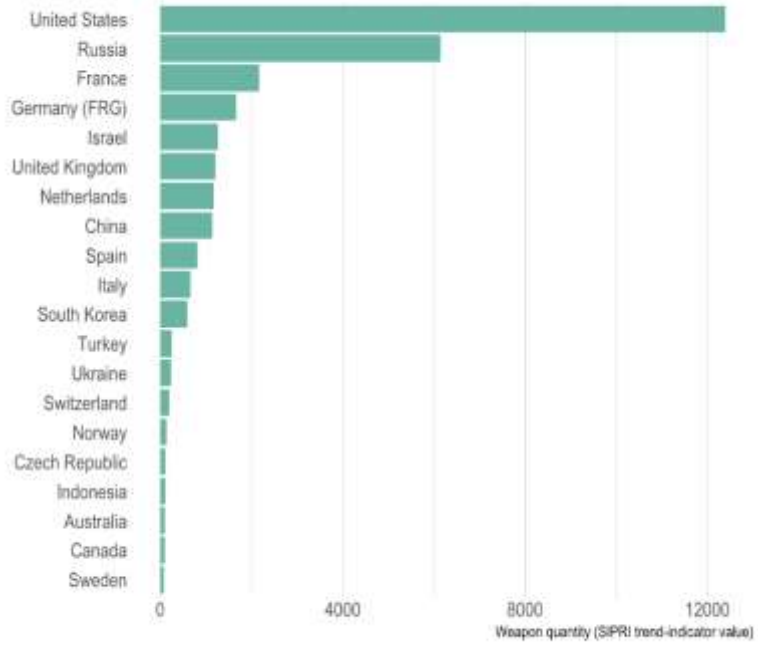
- Shows **proportion of each category** as part of a circle
- Useful for showing **percentage contribution**

Bar Plot

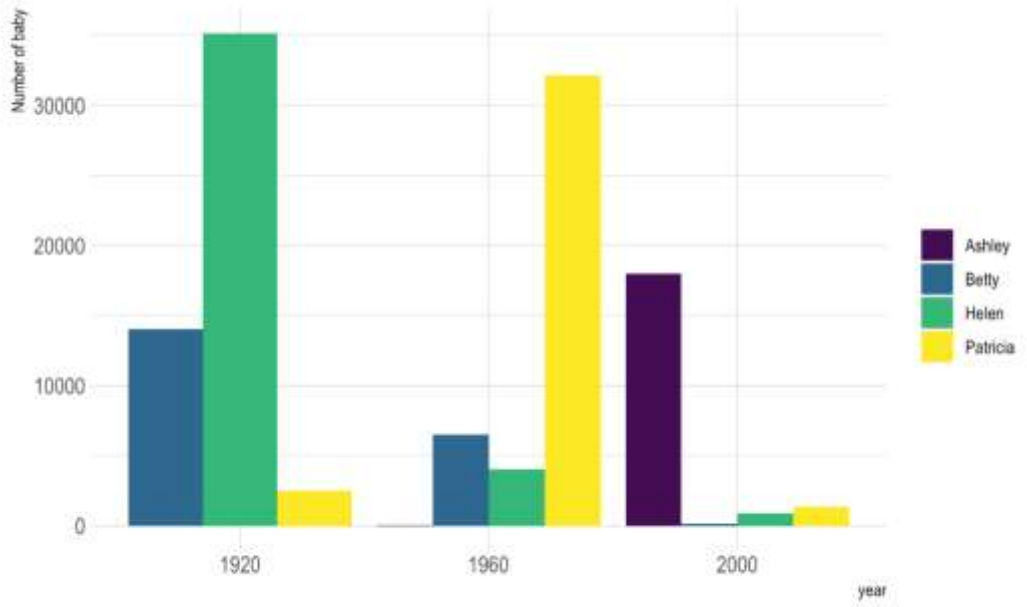
A barplot (or barchart) is one of the most common types of graphic. It shows the relationship between a numeric and a categoric variable. Each entity of the categoric variable is represented as a bar. The size of the bar represents its numeric value.

Type	Description	Example Use
Vertical Bar Chart	Bars stand upright; common for categorical data.	Comparing favorite fruits, countries, products.
Horizontal Bar Chart	Bars go sideways; useful when category names are long.	Comparing survey responses or countries.
Grouped (Clustered) Bar Chart	Multiple bars per category (to compare subgroups).	Comparing male vs female preferences per category.
Stacked Bar Chart	Bars divided into segments showing parts of a whole.	Showing category breakdown within totals.

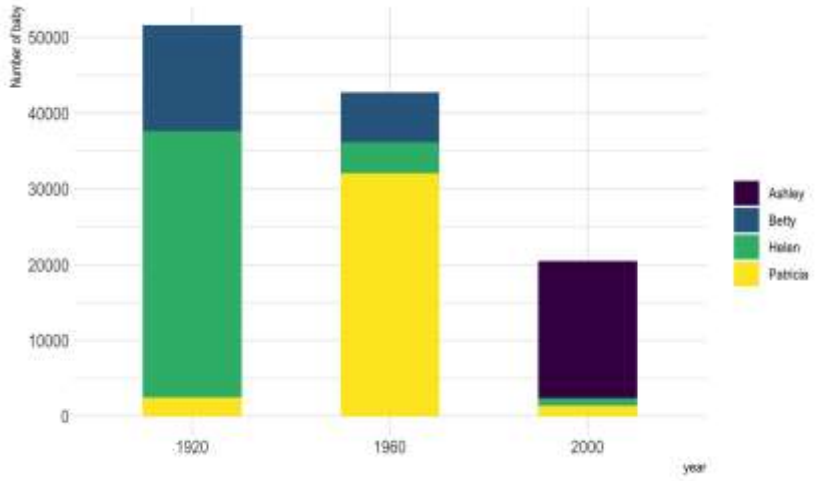
Here is an example showing the quantity of weapons exported by the top 20 largest exporters in 2017



A barplot can also display values for several levels of grouping. In the following graphic, the number of given baby name is provided.



Instead of putting the bars one beside each other it is possible to stack them, resulting in a stacked barplot:



Binomial distribution

The **Binomial Distribution** is a **probability distribution** that describes the number of **successes** in a fixed number of **independent trials**, where each trial has **only two possible outcomes**:

- **Success** (e.g., heads, yes, pass)
- **Failure** (e.g., tails, no, fail)

Example:

- Flipping a coin 5 times → heads = success, tails = failure
- Rolling a die and counting only "6" as success

Analogy:

Imagine a basketball player taking **10 shots**:

- Each shot = independent trial
- Success = scoring a basket
- Failure = missing
- Binomial distribution tells you the **probability of scoring exactly 7 baskets**, or 5 baskets, etc.

Formula:

The binomial distribution formula is:

$$\mathbf{b(x; n, P) = {}_nC_x * P^x * (1 - P)^{n-x}}$$

Where:

- b = binomial probability
- ${}_nC_x$ = combinations formula ${}_nC_x = n! / (x!(n - x)!)$

- x = total number of "successes"
- P = probability of a success on a single attempt
- n = number of attempts or trials.

Steps for computation;

- ***Step 1:*** *Find the number of trials and assign it as 'n'.*
- ***Step 2:*** *Find the probability of success in each trial and assign it as 'p'*
- ***Step 3:*** *Find the probability of failure and assign it as q where q = 1-p*
- ***Step 4:*** *Find the random variable X = r for which we have to calculate the binomial distribution*
- ***Step 5:*** *Calculate the probability of Binomial Distribution for X = r using the Binomial Distribution Formula.*

Numerical:

2. **A coin is tossed 10 times. What is the probability of getting exactly 6 heads?**

   Solution:

   We have formula: $b(x; n, P) - {}_nC_x * P^x * (1 - P)^{n-x}$

   The number of trials (n) is 10

   The odds of success ("tossing a heads") is 0.5 (So 1-p = 0.5) x = 6

   $P(x=6) = {}_{10}C_6 * 0.5^6 * 0.5^4$

   $= 210 * 0.015625 * 0.0625$

   $= 0.205078125$

Assignment:

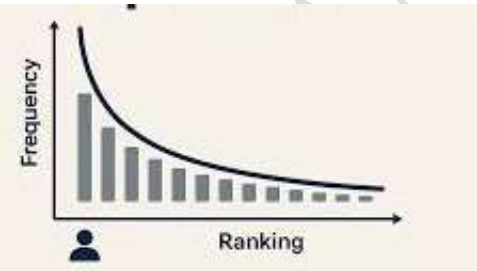
1. A die is thrown 6 times and if getting an even number is a success what is the probability of getting (i) 4 Successes (ii) No success
2. 80% of people who purchase pet insurance are women. If 9 pet insurance owners are randomly selected, find the probability that exactly 6 are women.
3. 60% of people who purchase sports cars are men. If 10 sports car owners are randomly selected, find the probability that exactly 7 are men.
4. If we toss a coin 20 times and getting head is the success then what is the mean of success and variance of the distribution and standard deviation?

Zipf's law

- Zipf's law is an empirical formula discovered by George Zipf in 1930s.
- Zip's law describes the relationship between the frequency of words in language corpus and their rank in a frequency sorted list.
- Zipf's Law describes a pattern in ranked data, where:
    - The frequency of an item is inversely proportional to its rank in a list.
    - Mathematically:

$$f(r) \propto \frac{1}{r^s}$$

    - Where:
        - f(r) = frequency of the item
        - r= rank of the item (1 = most frequent, 2 = second most frequent, …)
        - s = parameter (usually ≈ 1)



Simple Numerical Example

Suppose we rank 5 words by frequency in a text:

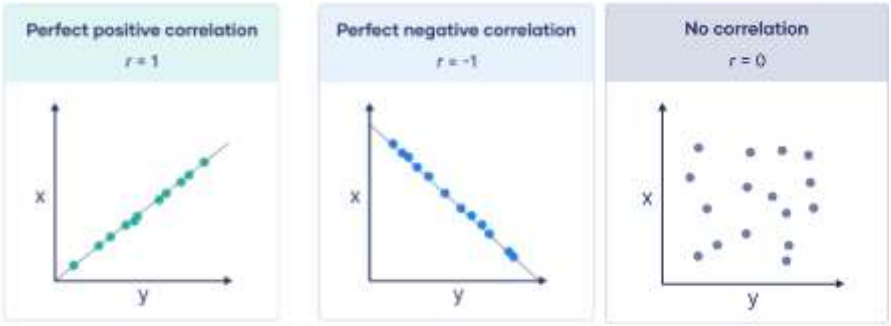
Rank (r)	Frequency f(r) ≈ 1/r
1	1 / 1 = 1.0
2	1 / 2 = 0.5
3	1 / 3 ≈ 0.33
4	1 / 4 = 0.25
5	1 / 5 = 0.2

Interpretation:

- Word ranked 1 → occurs **most often**
- Word ranked 5 → occurs **rarely**
- Shows the **"few frequent, many rare"** pattern

3. Correlation Analysis

Correlation analysis studies **how strongly two variables are related** and **how one changes with another**.



- **Positive Correlation:**
  - Positive correlation indicates that two variables have a direct relationship. As one variable increases, the other variable also increases.
  - For example, there is a positive correlation between height and weight. As people get taller, they also tend to weigh more.
- **Negative Correlation:**
  - Negative correlation indicates that two variables have an inverse relationship. As one variable increases, the other variable decreases.
  - For example, there is a negative correlation between price and demand. As the price of a product increases, the demand for that product decreases.
- **Zero Correlation:**
  - Zero correlation indicates that there is no relationship between two variables. The changes in one variable do not affect the other variable.
  - For example, there is zero correlation between shoe size and intelligence.

a) Pearson Correlation: correlation between numeric variables

- Pearson correlation coefficient (**r**) measures the **strength and direction of a linear relationship** between two numerical variables.
  - **Range**: −1 to +1
  - **r = +1** → perfect increasing linear relationship
  - **r = −1** → perfect decreasing linear relationship

- **r = 0** → no linear relationship

- **Formula:**

$$r = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

Analogy: Imagine two people walking:

- If they always walk **side by side**, that's **+1 correlation**.
- If one walks forward while the other walks backward, that's **–1 correlation**.
- If one walks randomly without matching the other, that's **0 correlation**.

Real-Life Applications (Data Science)

- **Marketing:** More ads → more sales (positive correlation).
- **Health:** More exercise → lower weight (negative correlation).
- **Finance:** Stock prices and interest rates.
- **Machine Learning:**
  - Feature selection (remove highly correlated features).
  - Detecting multicollinearity.

Numerical Example

Dataset:

X = number of study hours = [2, 3, 5, 6]

Y = marks = [50, 60, 80, 85]

Step 1: Compute Means

$$\bar{X} = \frac{2+3+5+6}{4} = \frac{16}{4} = 4$$

$$\bar{Y} = \frac{50+60+80+85}{4} = \frac{275}{4} = 68.75$$

Step 2: Computer Covariance

$$\text{cov}(X, Y) = \frac{(2-4)(50-68.75) + (3-4)(60-68.75) + (5-4)(80-68.75) + (6-4)(85-68.75)}{n-1}$$

$$\text{cov} = 62.5$$

Step 3: Standard deviations

$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}} = \frac{10}{3} = 3.3333$$

$$\sigma_X = \sqrt{3.3333} = 1.8257 \approx 1.825$$

$\sigma_x = 1.825$

$\sigma_y = 16.77$

Step 4: Compute Correlation

$$r = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

$$r = \frac{62.5}{(1.825)(16.77)} = 0.95$$

Step 5: Interpretation

Strong positive correlation

Python code:

```
import numpy as np
from scipy.stats import pearsonr
X = np.array([2, 3, 5, 6])
Y = np.array([50, 60, 80, 85])
corr, p_value = pearsonr(X, Y)
print("Correlation:", corr)
print("p-value:", p_value)
```

Assignment:

Question: Marks obtained by 5 students in algebra and trigonometry as given below:

Algebra	15	16	12	10	8
Trigonometry	18	11	10	20	17

Calculate the Pearson correlation coefficient.

Ans: -0.424

b) Cross-tabulation: Correlation between categorical variables

Cross-tabulation (also called **contingency table**) is a **tabular method to show the frequency distribution of two or more categorical variables**.

Cross-tabulation is a special technique to organize data in table format which facilitates a clear and concise representation of the relationships between categorical variables.

It helps to:

- Identify **relationships or association** between categorical variables
- Test whether variables are **independent or correlated**

The most common **statistical test** associated with cross-tabulation is **Chi-square test of independence**.

Analogy

Imagine a **classroom survey**:

- Variable 1: Gender → Male, Female
- Variable 2: Favorite subject → Math, Science

A cross-tabulation shows how many **males like Math, males like Science**, etc.

- If most males like Math and most females like Science, there is a **relationship**.
- If numbers are similar across categories, variables are **independent**.

Numerical:

Suppose we collect data from **20 students** about their **Study Method** and **Exam Result**. Here, G: Group Study, S:self Study, P: Pass and F:Fail. And the dataset is given below. Create a cross tabulation.

Student	Study Method	Result
1	G	P
2	S	P
3	S	F
4	G	P

Student	Study Method	Result
5	G	P
6	S	P
7	G	F
8	S	P
9	S	P
10	G	F
11	G	P
12	S	F
13	S	P
14	G	P
15	G	P
16	S	P
17	S	F
18	G	P
19	S	P
20	S	P

Solution:

Step 1: Count Frequencies: Count how many students fall into each combination:

Study Method	Pass	Fail
Group Study (G)	7	2
Self Study (S)	8	3

Step 2: Create the Cross-Tabulation Table

Study Method	Pass	Fail	Total
Group Study (G)	7	2	9

Study Method	Pass	Fail	Total
Self Study (S)	8	3	11
Total	**15**	**5**	**20**

Step 3 (Optional): Convert to Percentages

Row Percentages (Effect of method on result)

- Group Study: Pass = 7/9 = 77.8%; Fail = 2/9 = 22.2%
- Self Study: Pass = 8/11 = 72.7%; Fail = 3/11 = 27.3%

Column Percentages (Distribution of each result across methods)

- Pass: G = 7/15 = 46.7%; S = 8/15 = 53.3%
- Fail: G = 2/5 = 40%; S = 3/5 = 60%

Interpretation

- Both study methods lead to mostly passing results.
- **Group Study pass rate = 77.8%**
- **Self Study pass rate = 72.7%**

This kind of cross-tab helps analyze **relationship between two categorical variables**, and later can be used with **Chi-square test** for association.

Real-Life Applications (Data Science)

1. **Marketing analytics:**
   - Does age group affect choice of product?
   - Are males more likely to buy Brand A than females?
2. **Healthcare:**
   - Is disease type associated with smoking status?
3. **Social science surveys:**
   - Does education level influence voting preference?
4. **Fraud detection:**
   - Does transaction type vary with location category?

Used to check:

- **Association**
- **Independence**
- **Trend patterns**

c) Assessing correlation between numeric and categorical variables

When one variable is **numeric** (e.g., income, marks, age) and the other is **categorical** (e.g., gender, region, study method), we cannot directly use **Pearson correlation**, because Pearson requires *both* variables to be numeric.

Instead, we use **statistical methods that compare group means** or **explain variation in the numeric variable due to categories**.

Below are three common methods.

i) Using Difference of Means (Two Categories)

If the categorical variable has **two groups** (e.g., Male/Female, Urban/Rural), we compare the **mean values** of the numeric variable.

Example

Categorical variable: **Study Method** = {Group Study, Self Study}
Numeric variable: **Marks**

Study Method	**Marks**
Group Study	72, 80, 65, 78
Self Study	60, 55, 58, 62

Compute Means

- Mean (Group Study) = (72+80+65+78)/4 = **73.75**
- Mean (Self Study) = (60+55+58+62)/4 = **58.75**

Interpretation

The difference in average marks (**73.75 vs 58.75**) suggests the numeric variable (marks) **depends on** the categorical variable (study method). A **t-test** can be used to check if this difference is statistically significant.

ii) Using ANOVA (More Than Two Categories)

If the categorical variable has **3 or more categories** (e.g., Regions = East, West, North), use **Analysis of Variance (ANOVA)**. ANOVA tests whether the **group means of the numeric variable are significantly different**.

Example

Categorical variable: **Region** = {East, West, North}

Numeric variable: **Income**

Suppose the average incomes are:

- East = Rs. 20,000
- West = Rs. 22,500
- North = Rs. 18,000

If ANOVA shows a significant result, it means: **Income depends on (is correlated with) region.**

iii) Using Point-Biserial Correlation (Special Case)

If the categorical variable has **exactly two categories**, we can convert it into numeric labels {0, 1} and use **point-biserial correlation**.

Example

Gender: Male = 0, Female = 1

Marks: numeric values

The formula for point-biserial correlation shows how strongly the numeric variable varies with the two-category variable.

This method is mathematically equivalent to comparing group means.

4. Statistical significance

Statistical significance helps us decide whether an observed pattern in data is **real** or just due to **random chance**.

We use different statistical tests depending on the type of data and the question. The results of these tests are often expressed through:

- **p-value**
- **t-value**
- **Chi-squared (χ^2)**

Each plays a different role.

a) P-value

The **p-value** is the probability of getting a result **at least as extreme as your observed result**, **if the null hypothesis is true**.

- Low p-value → result is unlikely to be due to chance
- High p-value → result could easily be due to chance

Interpretation

- **$p < 0.05$** → statistically significant
- **$p \geq 0.05$** → not significant (evidence not strong)

Simple Example

Suppose you compare the mean marks between two groups (boys vs girls), and get:

$p=0.03$

This means:
There is only a 3% chance that the observed difference happened by random chance.
So we say the difference is **statistically significant**.

b) T-value

The **t-value** comes from the **t-test**, used when comparing **means of two groups**.
For example:

- Does a new teaching method increase average test scores?
- Is there a difference between two groups?

Interpretation

- **Large t-value** → groups are very different
- **Small t-value** → groups are similar

Simple Numerical Example

Group A marks: 60, 62, 63

Group B marks: 70, 72, 71

If the computed **t-value = 4.2**, and p-value < 0.05, then:

The difference between A and B is statistically significant.

c) Chi-squared

Chi-squared is used for **categorical data**, especially:

- Cross-tabulation tables
- Testing association between variables (gender vs pass/fail, region vs preference)

Definition

The χ^2 statistic measures the difference between:

- **Observed frequencies** (actual data), and
- **Expected frequencies** (what we would get if variables were NOT related)

χ^2 Formula

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

Simple Example

Study Method	**Pass**	**Fail**	**Total**
Group Study	20	10	30
Self Study	10	20	30
Total	30	30	60

Step 1: Compute Expected Values for Each Cell

$$E = \frac{(\text{Row Total}) \times (\text{Column Total})}{\text{Grand Total}}$$

Cell 1: Group Study – Pass

$$E = \frac{30 \times 30}{60} = 15$$

Cell 2: Group Study – Fail

$$E = \frac{30 \times 30}{60} = 15$$

Cell 3: Self Study – Pass

$$E = \frac{30 \times 30}{60} = 15$$

Cell 4: Self Study – Fail

$$E = \frac{30 \times 30}{60} = 15$$

So expected frequencies:

Study Method	Pass (E)	Fail (E)
Group Study	15	15
Self Study	15	15

Step 2: Compute χ^2 for Each Cell

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Cell 1: Group Study – Pass

$$\frac{(20-15)^2}{15} = \frac{25}{15} = 1.67$$

Cell 2: Group Study – Fail

$$\frac{(10-15)^2}{15} = \frac{25}{15} = 1.67$$

Cell 3: Self Study – Pass

$$\frac{(10-15)^2}{15} = \frac{25}{15} = 1.67$$

Cell 4: Self Study – Fail

$$\frac{(20-15)^2}{15} = \frac{25}{15} = 1.67$$

Step 3: Sum All χ^2 Components

$\chi 2$ = 1.67+1.67+1.67+1.67=6.68

Step 4: Degrees of Freedom

df = (r−1)(c−1)=(2−1)(2−1)=1

Step 5: Interpretation

Using a chi-square table:

- Critical value for df = 1 at $\alpha = 0.05$ is **3.84**
- Our value **6.68 > 3.84**

So:

Result is statistically significant. Study Method and Pass/Fail outcome are associated. ***This means students using the*** *Group Study* ***method have different results compared to*** *Self Study.*

***