FirstHope Notes | B.Pharm Semester 8

Regression

Regression in Biostatistics

Introduction

- Regression analysis is a statistical method used to study the relationship between:
  - one **dependent variable**, and
  - one or more **independent variables**.
- In **biostatistics**, regression is very important because it helps researchers:
  - understand medical and biological data,
  - study the effect of risk factors on health,
  - predict health outcomes, and
  - make better decisions in medicine and public health.

Properties (Assumptions) of Regression

For regression results to be valid, the following conditions should usually be satisfied:

- **Linearity:** The relationship between the variables should be linear.
- **Additivity:** The combined effect of more than one independent variable is added together.
- **Independence:** Each observation should be independent of the others.
- **Homoscedasticity:** The spread (variance) of errors should be the same for all values of the independent variables.
- **Normality:** The errors (residuals) should be approximately normally distributed.

Methods to Study Regression

Regression models can be estimated using different methods:

- **Least Squares Method:** Finds the line that minimizes the total squared errors.
- **Maximum Likelihood Estimation (MLE):** Finds the values of the parameters that make the observed data most likely.
- **Bayesian Methods:** Use previous knowledge (prior information) and update it with data.
- **Robust Regression:** Works better when there are outliers or when assumptions are not fully satisfied.

Types of Regression

Different regression models are used in biostatistics depending on the type of data:

- **Simple Linear Regression:** One independent variable and one dependent variable.
- **Multiple Linear Regression:** Two or more independent variables and one dependent variable.
- **Logistic Regression:** Used when the outcome is binary (for example: disease / no disease).
- **Poisson Regression:** Used for count data.
- **Cox Proportional Hazards Regression:** Used in survival or time-to-event analysis.
- **Non-linear Regression:** Used when the relationship is not linear.
- **Polynomial Regression:** Uses polynomial terms (square, cube, etc.).
- **Ridge and Lasso Regression:** Used to prevent overfitting by shrinking large coefficients.

Regression Equations

For X on Y:

$$X = a + bY$$

- where:
  - a is the intercept (value of X when Y = 0),
  - b is the slope (rate of change of X for each unit change in Y).

For Y on X:

$$Y = a + bX$$

- where:
  - a is the intercept (value of Y when X = 0),
  - b is the slope (rate of change of Y for each unit change in X).

Regression Line

- A regression line is a straight line that best fits the data points.
- It shows the relationship between the independent and dependent variables.
- It is used to predict the value of the dependent variable.

Applications in Biostatistics

- Regression analysis is widely used in:
  - epidemiology,
  - pharmacology, and
  - clinical trials.
- It is mainly used for:
  - **Prediction**: Estimating outcomes based on input variables (e.g., predicting blood pressure from body weight).

- **Understanding Relationships**: Evaluating how variables are associated (positive or negative relationship).
- **Assessing Model Fit**: Metrics like R-squared and Adjusted R-squared help assess how well the model explains the variability in the outcome.

Example 1: Educational Context

Dataset

Hours Studied (X)	Exam Score (Y)
1	50
2	55
3	65
4	70
5	80
6	85

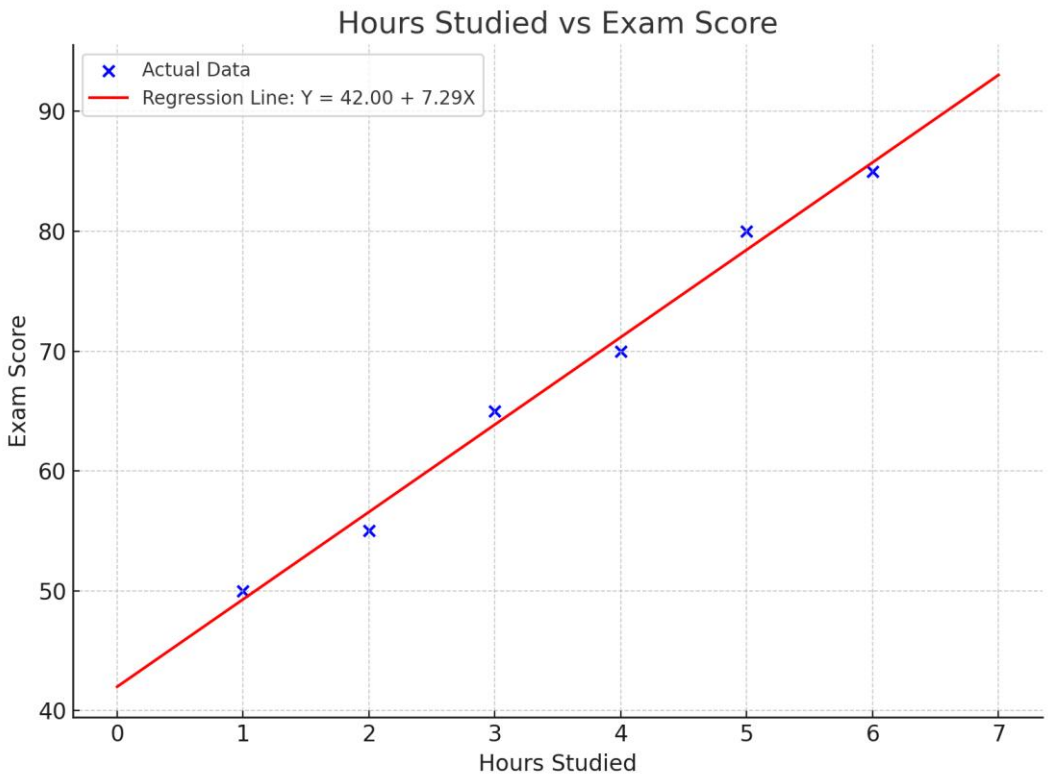
Regression Equation

- After applying the least squares method:

$$Y = 42.00 + 7.29X$$

Interpretation:

- **Intercept (42.00)**: The baseline score without studying is 42.
- **Slope (7.29)**: Each additional hour of study increases the score by ~7.29 points.



Curve Fitting Using Least Squares Method

Meaning of Curve Fitting

- Curve fitting means finding a **best-fit line or curve** that shows the trend of the given data.
- The **least squares method** is used to find this best-fit line.
- It works by **minimizing the sum of the squared differences** between:
  - the observed values, and
  - the predicted values.

Linear Models

Two simple linear models are commonly used:

- **y = a + bx**
  - Used to **predict y from x**
- **x = a + by**
  - Used to **predict x from y**
  - This is called the **reverse regression**

Example 2: Pharmaceutical Application

Dataset

Dosage (mg) (x)	Concentration (mg/L) (y)
10	1.2
20	2.5
30	3.1
40	4.7
50	5.5

Fitting the model: y = a + bx

Step 1: Calculate the means

$$\bar{x} = \frac{10 + 20 + 30 + 40 + 50}{5} = 30$$

$$\bar{y} = \frac{1.2 + 2.5 + 3.1 + 4.7 + 5.5}{5} = 3.4$$

Step 2: Calculate slope (b)

$$b = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$b \approx 0.112$$

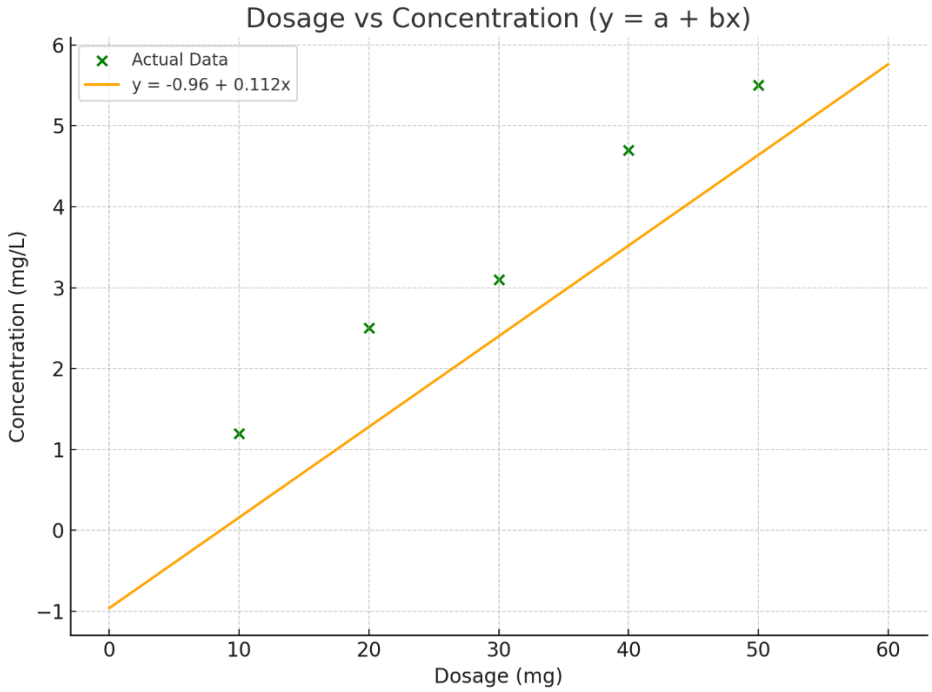
Step 3: Calculate intercept (a)

$$a = \bar{y} - b\bar{x}$$

$$a = 3.4 - (0.112 \times 30) = -0.96$$

Regression equation (y on x)

$$\boxed{y = -0.96 + 0.112x}$$



Fitting the model: x = a + by

Step 1: Use the same means

$$\bar{x} = 30, \bar{y} = 3.4$$

Step 2: Calculate slope (b)

$$b = \frac{\sum(y_i - \bar{y})(x_i - \bar{x})}{\sum(y_i - \bar{y})^2}$$

$$b \approx 9.12$$

Step 3: Calculate intercept (a)

$$a = \bar{x} - b\bar{y}$$

$$a = 30 - (9.12 \times 3.4) = -1.01$$

Regression equation (x on y)

$$\boxed{x = -1.01 + 9.12y}$$

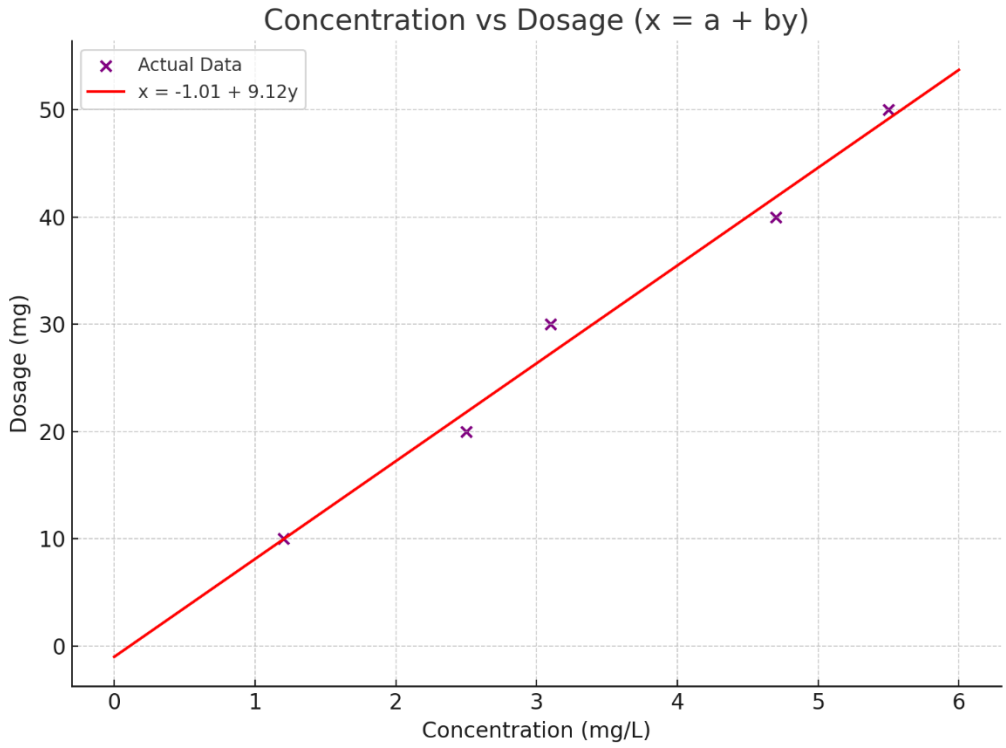


Table Summary: Predictions from Both Models

Dosage (mg)	Concentration (mg/L)	Predicted concentration (from y = −0.96 + 0.112x)	Predicted dosage (from x = −1.01 + 9.12y)
10	1.2	0.16	9.52
20	2.5	1.28	21.79
30	3.1	2.40	27.96
40	4.7	3.52	43.85
50	5.5	4.64	49.66

Important Note (in simple words)

- The model **y = −0.96 + 0.112x** is used to:
  - predict **concentration** from **dosage**.
- The model **x = −1.01 + 9.12y** is used to:
  - predict **dosage** from **concentration**.
- These two regression lines are **not the same**, because each one minimizes error in a different direction.

Multiple Regression in Biostatistics

Introduction

- Multiple regression is an extension of simple linear regression.

- It is used when:
  - there is **one dependent variable**, and
  - **two or more independent variables**.
- In biostatistics and pharmaceutical research, many factors can affect an outcome at the same time.
- Multiple regression helps to study the combined effect of:
  - dosage,
  - age,
  - weight,
  - and other biological or demographic variables.

Multiple Regression Model

- The general form of the multiple regression equation is:

$$y = a + b_1x_1 + b_2x_2 + \cdots + b_kx_k$$

- Where:
  - **y** = dependent variable
    - (for example: drug concentration)
  - $\mathbf{x_1, x_2, \ldots, x_k}$ = independent variables
    - (for example: dosage, age, weight)
  - **a** = intercept
  - $\mathbf{b_1, b_2, \ldots, b_k}$ = regression coefficients for each independent variable

Example Dataset

- Let's consider a dataset from a pharmaceutical study evaluating the effect of dosage, age, and weight on the concentration of a drug in blood.

Dosage (mg)	Age (years)	Weight (kg)	Concentration (mg/L)
10	25	70	1.2
20	30	80	2.5
30	35	90	3.1
40	40	85	4.7
50	50	95	5.5

Fitted Multiple Regression Model

➢ The fitted multiple regression model (based on the data above) is:

$$\hat{y} = 0.159 + 0.108 \times \text{Dosage} + 0.003\ \text{Age} + 0.013 \times \text{Weight}$$

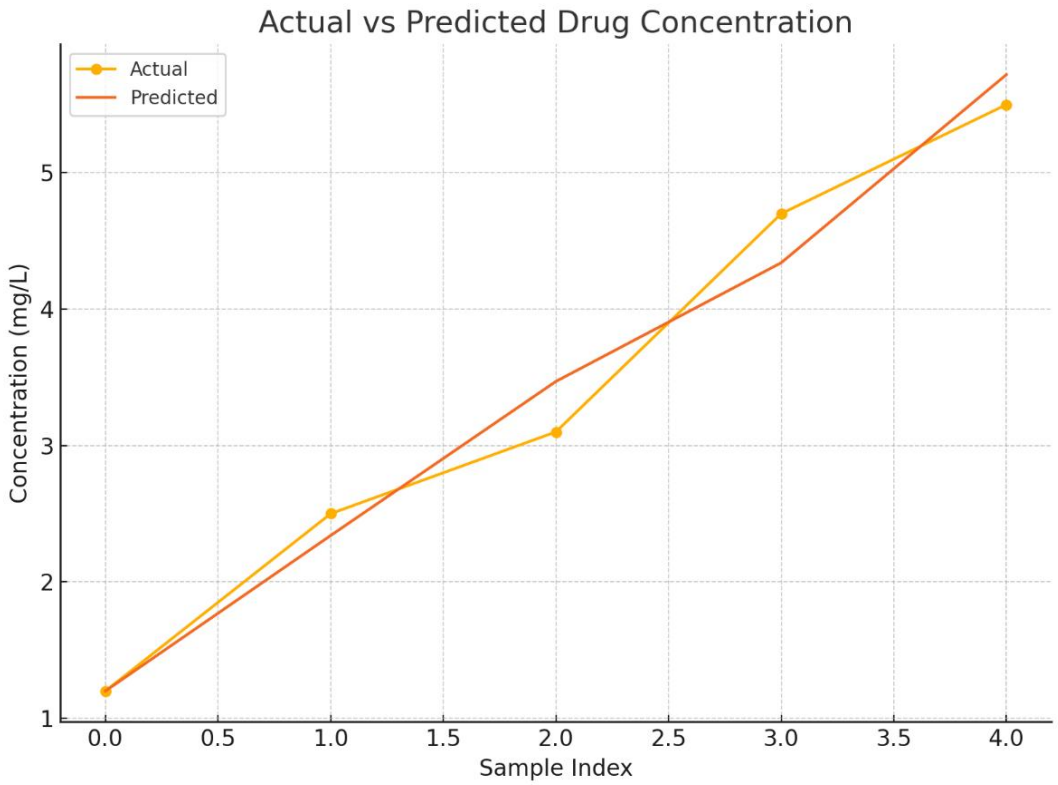
- Where:
  - $y^{\hat{y}}$ is the **predicted concentration**
  - Coefficients represent the **change in concentration** for each unit increase in the corresponding variable, holding others constant

Predicted vs. Observed Values

Dosage (mg)	Age (years)	Weight (kg)	Actual Concentration (mg/L)	Predicted Concentration (mg/L)
10	25	70	1.2	1.20
20	30	80	2.5	2.34
30	35	90	3.1	3.47
40	40	85	4.7	4.34
50	50	95	5.5	5.72

Interpretation

- Drug concentration generally increases when:
  - dosage increases,
  - age increases, and
  - weight increases.
- The predicted concentrations are very close to the actual concentrations.
- This shows that the multiple regression model provides a **good fit** to the data.



Standard Error of the Regression (SER)

Definition

- The **standard error of the regression (SER)** tells us:
  - how far, on average, the observed values are from the regression line.
- In simple words:
  - it measures the **typical prediction error** of the model.

Formula

$$\text{SER} = \sqrt{\frac{1}{N-k-1}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

- Where:
  - y_i = observed (actual) value
  - $\hat{y}_i$ = predicted value
  - **N** = number of observations
  - **k** = number of independent variables

Graphical Representation

- To visualize the model:
  - **X-axis**: Fitted or predicted concentrations
  - **Y-axis**: Actual concentrations
  - **Red Line**: Ideal fit line where predicted = actual
  - **Dots**: Actual values
  - The closer the dots are to the red line, the **better the model fit**.

Pharmaceutical Example – Simple Linear Regression

Scenario

➢ Researchers are investigating how **drug dosage (mg)** affects **symptom reduction** on a standardized scale.

Dataset

Dosage (mg)	Symptom Reduction
5	3
10	7
15	8

20	10
25	15
30	18
35	21
40	22
45	24
50	28

Fitted Linear Model

$$\hat{y} = 1.69 + 0.52x$$

- (Where y is symptom reduction and xxx is dosage in mg)

Predictions and Residuals

Dosage (mg)	Actual Reduction	Predicted Reduction	Residual (Actual – Predicted)
5	3	3.33	-0.33
10	7	6.05	0.95
15	8	8.78	-0.78
20	10	11.51	-1.51
25	15	14.24	0.76
30	18	16.96	1.04
35	21	19.69	1.31
40	22	22.42	-0.42
45	24	25.15	-1.15
50	28	27.87	0.13

Standard Error of Regression (SER)

$$\text{SER} \approx 1.05$$

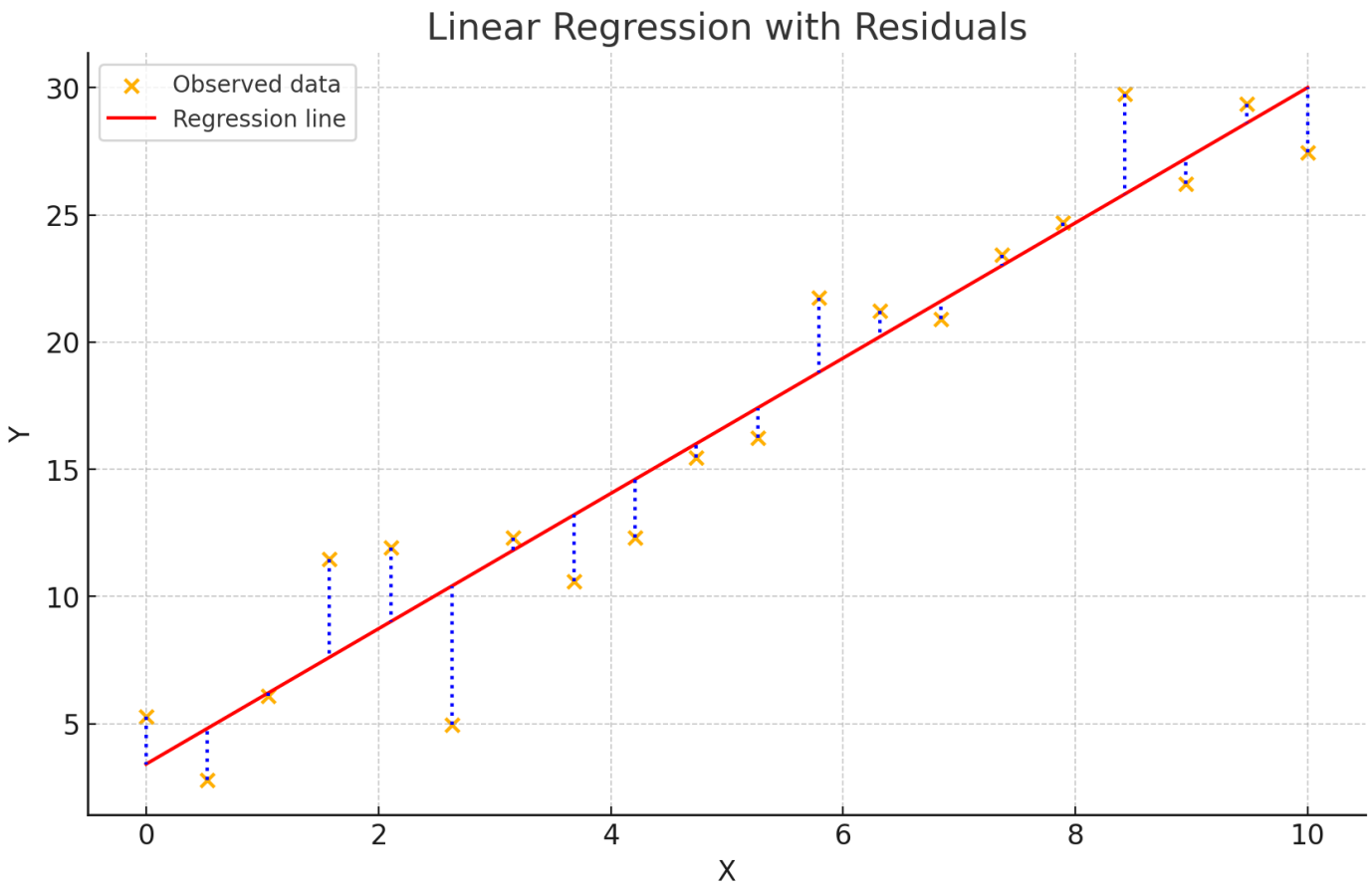
Interpretation

- The typical error between actual and predicted symptom reduction is about **1.05 units**, indicating a **reasonably accurate model**.

Graphical Explanation

- **Blue Dots**: Actual data points
- **Red Line**: Regression line (predicted values)
- **Vertical Lines**: Residuals (errors between actual and predicted)

This plot visually represents how well the model fits the data, showing the **magnitude of residuals** and whether the **model under- or overestimates** at different dosages.



Probability

Probability in Biostatistics

Definition of Probability

- Probability is a basic idea in biostatistics.
- It deals with **randomness and uncertainty** in health and biological studies.
- It helps us measure **how likely** an event is to happen in medical and biological research.

Basic Definition

- Probability is a number that shows the chance of an event occurring.
- Its value always lies between **0 and 1**.
- **0** → the event will not happen
- **1** → the event will definitely happen
- In biostatistics, probability helps researchers:
  - draw conclusions about a population using sample data.

Role of Probability in Biostatistics

- In biostatistics, probability is used to:
  - **Design Experiments**: Determine the sample size needed to achieve a certain level of confidence.
  - **Analyze Data**: Assess the likelihood of observing the data if a certain hypothesis is true.
  - **Make Decisions**: Make informed decisions based on the probability of various outcomes.

Types of Probability

1. **Theoretical Probability**
   - Based on the assumption of equally likely outcomes.
   - ***Example*:** Probability of drawing a specific genotype from a gene pool.
2. **Empirical Probability**
   - Based on observed data.
   - ***Example*:** Probability of a patient recovering from a disease based on clinical trial results.
3. **Subjective Probability**
   - Based on personal judgment or experience.
   - ***Example*:** An expert's estimate of a new drug's likelihood of success.

Probability in Medical Research

- Probability is crucial for:
  - **Epidemiological Studies**: Estimating the probability of disease occurrence in different populations.
  - **Clinical Trials**: Assessing treatment efficacy and side effects.
  - **Risk Assessment**: Evaluating the probability of adverse events from specific treatments or conditions.

Calculation of Probability

1. Classical Approach

$$\text{Probability} = \frac{\text{Number of favourable outcomes}}{\text{Total number of possible outcomes}}$$

2. Frequency Approach

$$\text{Probability} = \frac{\text{Number of times an event occurs}}{\text{Total number of trials or observations}}$$

3. Bayesian Approach

- Incorporates **prior knowledge or experience** into the calculation of probability.

Applications in Biostatistics

- **Diagnostic Testing**: Calculating the probability of disease given test results (sensitivity/specificity).
- **Survival Analysis**: Estimating the probability of surviving a period after treatment.
- **Genetics**: Calculating the probability of inheriting traits or genetic disorders.

Probability Distributions

1. Binomial Distribution

- Models the number of successes in a fixed number of independent **Bernoulli trials** (yes/no outcomes).

2. Normal Distribution

- A continuous distribution that is symmetrical around the **mean**, often referred to as the **bell curve**.

3. Poisson Distribution

- Models the number of events occurring within a fixed **interval of time or space**.

Example

- A clinical trial includes **100 patients**.
- **80 patients** show a positive response.

- The empirical probability of a positive response is:

$$\text{Probability} = \frac{80}{100} = 0.8$$

Interpretation (in simple words)

- The probability of a patient responding positively to the drug is **0.8**.
- This means:
  - about **80% of patients** are expected to benefit.
- This information helps doctors and researchers judge the **effectiveness of the drug** and make better decisions.

Binomial Distribution

Meaning (in simple words)

- The binomial distribution is a discrete probability distribution.
- **It is used when:**
  - we perform a fixed number of trials,
  - each trial has only two possible outcomes (success or failure),
  - and each trial has the same probability of success.

When can we use a binomial distribution?

- A binomial situation must satisfy:
  - The number of trials is fixed (**n**).
  - Each trial is independent.
  - Each trial has only two outcomes:
    - success
    - failure
  - The probability of success (**p**) is the same for every trial.

Formula

- The probability of getting exactly **k successes** in **n trials** is:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

- Where:
  - **n** = total number of trials

- **k** = number of successes
- **p** = probability of success in one trial
- **(1 – p)** = probability of failure in one trial
- $\binom{n}{k}$= binomial coefficient

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

Graphical idea of a binomial distribution

- The horizontal axis shows the **number of successes (k)**.
- The vertical axis shows the **probability** of each value of k.
- The bars show that the distribution is **discrete**.

Applications in Biostatistics

- The binomial distribution is used when we count how many times something happens in a fixed number of trials.
- Common uses include:
  - **Clinical trials:** number of patients who respond positively to a treatment.
  - **Quality control:** number of defective items in a batch.
  - **Surveys:** number of people who agree with a statement.

Example

- A new drug is tested on **10 patients**.
  - Probability of a positive response for one patient = **0.7**
  - We want to find the probability that **exactly 7 patients** respond positively.

Step 1: Identify the values

- $n = 10$
- $k = 7$
- $p = 0.7$

Step 2: Apply the binomial formula

$$P(X = 7) = \binom{10}{7}(0.7)^7(0.3)^3$$

$$P(X = 7) = \frac{10!}{7!\,3!}(0.7)^7(0.3)^3$$

$$P(X = 7) = 120 \times 0.0823543 \times 0.027$$

$$P(X = 7) \approx 0.2668$$

Final answer

- The probability that **exactly 7 out of 10 patients** respond positively is:

$$\boxed{P(X = 7) \approx 0.2668}$$

Interpretation (in simple words)

- There is about a **26.68% chance** that exactly 7 patients will show a positive response to the drug.

Normal Distribution

Definition

- The **normal distribution** is a **continuous probability distribution**.
- It has a **bell-shaped curve**.
- The curve is **symmetric** around the mean (**μ**).
- It is widely used to describe how real-life measurements are distributed in science and medicine.

Probability Density Function (PDF)

$$\mathbf{f}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- where:
  - μ is the mean
  - σ is the standard deviation
  - e is the base of the natural logarithm (approximately 2.71828)
  - π is Pi (approximately 3.14159)

Applications

- The normal distribution is used to model real-valued random variables in fields like:
  - Heights and weights of individuals
  - Blood pressure readings
  - Test scores
  - Measurement errors

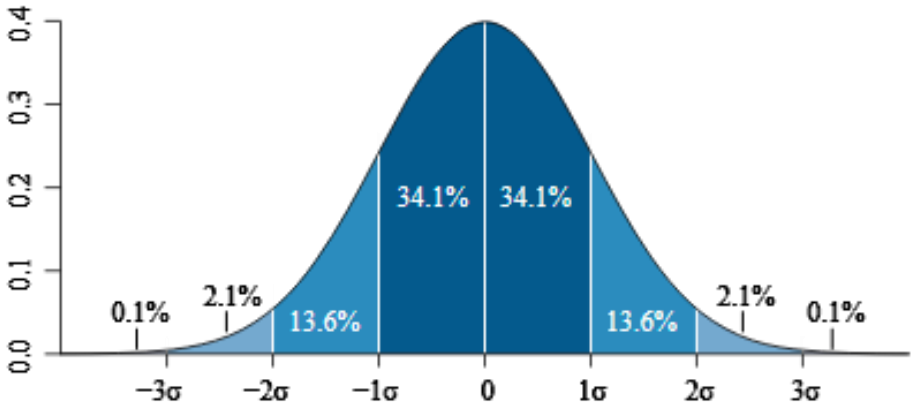
Properties

- **Symmetry**: Symmetric about the mean
- **Mean = Median = Mode**: All located at the center
- **Empirical Rule**:

- ~68% within 1 standard deviation
- ~95% within 2 standard deviations
- ~99.7% within 3 standard deviations

Normal Curve

- A graphical representation of the normal distribution:
  - Bell-shaped
  - Centered at μ
  - Spread defined by σ



Standard Normal Distribution

Definition

- The **standard normal distribution** is a special case of the normal distribution.
- It has:
  - **Mean (μ) = 0**
  - **Standard deviation (σ) = 1**
- The variable used is called **Z**.
- It is mainly used to find probabilities using **Z-tables**.

Probability Density Function (PDF)

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

- Where:
  - **z** = standard normal variable
  - **e** = base of the natural logarithm
  - **π** = pi

Why do we use the standard normal distribution?

- It allows us to convert any normal variable into a **common scale**.
- This makes it easy to calculate probabilities using a single table (Z-table).

Z-score formula

$$Z = \frac{X - \mu}{\sigma}$$

- Where:
  - **X** = actual value
  - **μ** = mean
  - **σ** = standard deviation

Example

Context

- Heights of adult men are normally distributed with:
  - Mean, **μ = 70 inches**
  - Standard deviation, **σ = 3 inches**

Find

- The probability that a randomly selected man is **between 68 and 72 inches** tall.

Step 1: Convert the values to Z-scores

$$Z = \frac{X - \mu}{\sigma}$$

- For **68 inches**:

$$Z = \frac{68 - 70}{3} = -0.67$$

- For **72 inches**:

$$Z = \frac{72 - 70}{3} = 0.67$$

Step 2: Find the probability

- We want:

$$P(-0.67 < Z < 0.67)$$

- From the Z-table:
  - $P(Z < 0.67) = 0.7486$
  - $P(Z < -0.67) = 0.2514$

- So,

$$P(-0.67 < Z < 0.67) = 0.7486 - 0.2514 = 0.4972$$

Final answer

$$\boxed{P(-0.67 < Z < 0.67) \approx 0.4972}$$

Conclusion (in simple words)

➢ There is about a **49.72% chance** that a randomly selected man is between **68 and 72 inches** tall.

Poisson's Distribution

Definition

- The **Poisson distribution** is a **discrete probability distribution**.
- It is used to find the probability of a certain **number of events** happening in a **fixed time or space interval**, when:
  - events occur **independently**, and
  - events occur at a **constant average rate**.

Formula

- The probability of observing exactly **k events** is:

$$P(X = k) = \frac{e^{-\lambda}\lambda^{k}}{k!}$$

- Where:
  - **λ (lambda)** = average number of events in the interval
  - **k** = actual number of events
  - **k!** = factorial of k
  - **e** = Euler's number (≈ 2.71828)

When is the Poisson distribution used?

- It is used when:
  - we count how many times something happens, and
  - the events are **rare or random**, and
  - the observation period is fixed (per day, per hour, per area, etc.).

Applications

- The Poisson distribution is commonly used to model:

- number of calls to a call centre per hour
- frequency of radioactive decay events
- number of emails received per day
- **rare medical or pharmaceutical events** (for example, side effects in patients)

Properties

- Type: **Discrete distribution**
- Parameter: **λ (average rate)**
- **Mean = λ**
- **Variance = λ**
- Best suited for **count data**

Pharmaceutical Example

Problem

- A pharmaceutical company observes, on average, **2 side effects per day** for a new drug.
- Find the probability of observing **exactly 3 side effects in one day**.

Given

$$\lambda = 2, k = 3$$

Step 1: Use the Poisson formula

$$P(X = 3) = \frac{e^{-2}2^3}{3!}$$

Step 2: Calculate

$$P(X = 3) = \frac{e^{-2} \times 8}{6}$$

$$P(X = 3) = \frac{0.1353 \times 8}{6}$$

$$P(X = 3) \approx 0.180$$

Final Answer

$$\boxed{P(X = 3) \approx 0.180}$$

Interpretation (in simple words)

- There is about an **18% chance** of observing **exactly 3 side effects in a day**.
- This shows how the Poisson distribution helps in:
  - modelling and assessing the risk of **rare medical events** in pharmaceutical and clinical studies.

Population and Sample in Biostatistics

Population

Definition

- A **population** is the **entire group** of individuals or observations that are the focus of a particular study.
- It includes **all members** of a defined group about whom or which data are intended to be collected or analyzed.

Characteristics

- **Size**:
  - **Finite** population: e.g., all patients in a hospital.
  - **Infinite** population: e.g., all possible measurements of a biological variable like blood pressure.
- **Parameters**:
  - Numerical values that describe characteristics of a population (e.g., **mean** μ, **variance** σ^2).

Example

- In a study investigating the **average blood pressure of adults** in a city, the **population** includes **all adults living in that city**.

Sample

Definition

- A **sample** is a **subset** of individuals or observations **selected from the population**.
- It is used to draw conclusions or make inferences about the **entire population**, without needing to examine everyone.

Characteristics

1. **Representativeness**:
   - A good sample accurately reflects the characteristics of the population.
2. **Size**:
   - The number of individuals or observations in the sample is called the **sample size**, denoted by n.
3. **Statistics**:
   - Numerical values calculated from a sample (e.g., **sample mean** $\bar{x}$, **sample variance** s^2).
   - These are used to **estimate population parameters**.

Example

- If we select **100 adults** from the city to measure their blood pressure, this group represents a **sample**.

Errors in Sampling

1. Sampling Error

- The **difference between a sample statistic** and the **actual population parameter** it is estimating.
- It arises because the sample is **only a subset** of the population and not the whole.

2. 2. Non-Sampling Error

- Errors that are **not related** to the act of sampling itself.
- These can result from:
  - **Measurement errors**
  - **Data processing errors**
  - **Bias** due to **non-random sampling methods**

Large Sample vs. Small Sample in Biostatistics

Large Sample

Definition

- A **large sample** means a sample size that is big enough to give **reliable and stable results**.
- There is no fixed rule, but in many cases:
  - **n > 30** is commonly treated as a large sample.
- Large samples allow the use of important results such as the **Central Limit Theorem (CLT)**.

Characteristics

A. Central Limit Theorem (CLT)

- As the sample size becomes large, the distribution of the sample mean becomes **approximately normal**, even if the population is not normal.

B. Greater precision

- Estimates are more accurate and stable.

C. Reduced sampling error

- Random variation due to sampling becomes smaller.

Advantages

- Gives **more accurate estimates** of population parameters.

- Results are more **robust** (less affected by outliers).
- Allows use of **normal approximation**, which simplifies many statistical tests.

Disadvantages

- **More time and cost** are needed to collect and process data.
- **Large datasets** can be harder to manage and analyse.

Example

- If a study wants to find the **average height of adult men in a city**, selecting **500 men** would be considered a **large sample**.
- This allows reliable conclusions about the whole population.

Small Sample

Definition

- A **small sample** means the number of observations is **limited**.
- It may not give very precise estimates of population parameters.
- In small samples, results such as the CLT **may not be valid**.

Characteristics

A. **Distributional assumptions**
   - Analyses often assume the data come from a **normal population**.

B. **Increased variability**
   - Results may change more from one sample to another.

C. **Use of t-distribution**
   - When the sample size is small (especially **n < 30**), the **t-distribution** is preferred instead of the normal distribution for estimating the mean.

Advantages

- **Cost-effective and time-saving**
- Easier to collect and analyse data.
- Useful in **early or pilot studies**.

Disadvantages

- **Less precise** estimates of parameters.
- **Higher risk of bias** if the sample is not representative.
- **Limited generalizability**
   - Results may not apply well to the whole population.

Example

- In a pilot study of a new drug, a sample of **15 patients** may be used.
- Because the sample is small:
  - analysis is done using the **t-distribution**, and
  - conclusions are made **carefully** due to higher variability and possible sampling error.

Hypothesis Testing

Hypothesis Testing

- **Hypothesis testing** is a statistical method used to make decisions about a **population** using **sample data**.
- It helps us decide whether there is **enough evidence** to support a claim about a population value (parameter).

Steps in Hypothesis Testing

1. Formulate Hypotheses

- **Null Hypothesis (H_0)**: Assumes no effect or no difference.
- **Alternative Hypothesis (H_1 or Ha)**: Assumes an effect or difference exists.

2. Choose Significance Level (α)

- Commonly used values: **0.05**, **0.01**, or **0.10**
- Represents the probability of **rejecting H_0 when it is true** (Type I error).

3. Collect Data

- Conduct the study or experiment and gather sample data.

4. Perform Statistical Test

- Use an appropriate test (e.g., **t-test**, **z-test**, **ANOVA**) to compute the **test statistic** and **p-value**.

5. Apply Decision Rule

- If **$p \leq \alpha$**: **Reject H_0**
- If **$p > \alpha$**: **Fail to reject H_0**

6. Draw Conclusion

- Interpret the results in the context of the study.
- State whether there is **sufficient evidence** to support H_1.

Null Hypothesis (H_0)

Definition

- The **null hypothesis** is a statement of no effect, no difference, or status quo.
- It serves as the baseline assumption in hypothesis testing.

Characteristics

- **Assumes no real effect**: Observed differences are due to random chance.
- **Default position**: Accepted unless sufficient evidence exists against it.
- **Includes equality**: Common forms include $=, \leq, or \ \geq$.

Pharmaceutical Example

- In a clinical trial testing a new drug:

$$H_0: \mu_{treatment} = \mu_{placebo}$$

- This suggests that the **mean outcome** for the treatment group is **equal** to the placebo group—**no effect** of the drug.

Alternative Hypothesis (H_1 or Ha)

Definition

- The alternative hypothesis is a statement that there is an effect or a difference.
- It represents what the researcher aims to prove.

Characteristics

- **Contradicts H_0**: Suggests that observed effects are **real**, not due to chance.
- **Directionality**:
  - **One-tailed**: Tests for a difference in a **specific direction** (e.g., greater than).
  - **Two-tailed**: Tests for a difference in **either direction**.
- **Uses inequality**: Common forms include $\neq, =, <, or \ >$.

Pharmaceutical Example

- Continuing from the drug trial:

$$H_1: \mu_{treatment} \neq \mu_{placebo}$$

- This suggests that the treatment group has a **different average outcome** from the placebo group—**indicating an effect** of the new drug.

Sampling

Definition

- **Sampling** is the process of selecting a **small group (subset)** from a large **population**.
- We study this small group to **understand and estimate** the characteristics of the whole population.
- Sampling is widely used in:
    - biostatistics,
    - social sciences,
    - medical research,
    - and marketing studies.

Advantages of Sampling

1. **Cost-Effective**: Cheaper than studying the whole population.
2. **Time-Saving**: Faster to study a sample.
3. **Feasible**: Easier to manage smaller groups.
4. **Detailed Analysis**: Allows for more in-depth study.
5. **Reduced Workload**: Smaller data sets are easier to analyze.

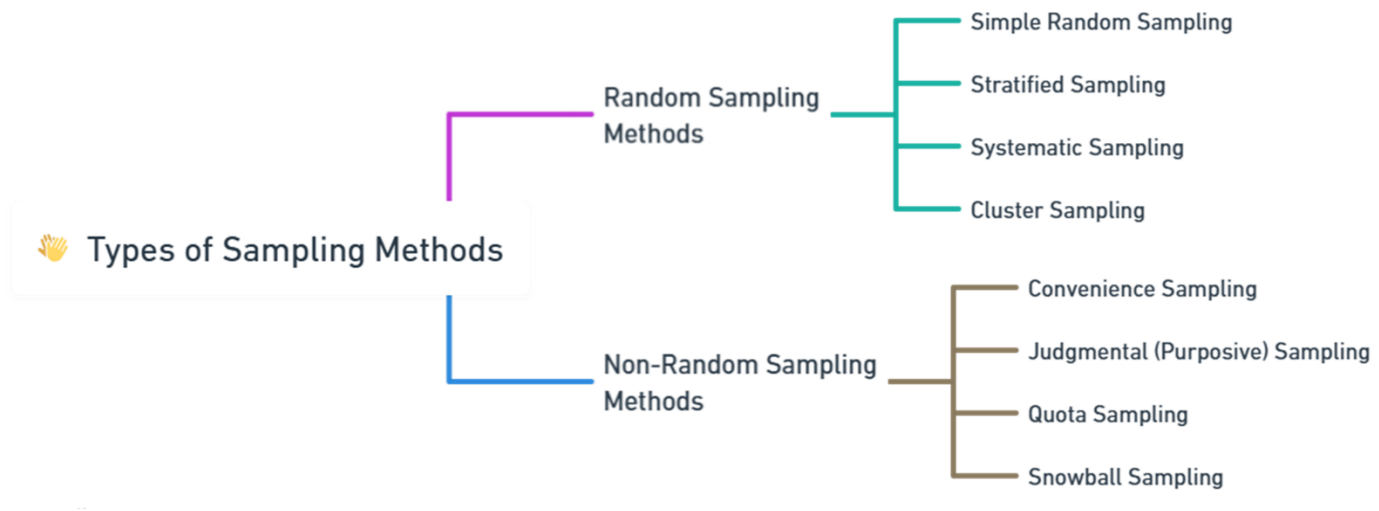
Disadvantages of Sampling

1. **Sampling Error**: The sample may not perfectly represent the population.
2. **Bias**: Poor sampling methods can lead to biased results.
3. **Variability**: Results can differ between samples.
4. **Non-Sampling Error**: Mistakes in data collection or analysis can occur.

Essence of Sampling

- The main idea of sampling is to study a population in a **practical and efficient way**.
    - **Efficiency** – studying a sample is quicker.
    - **Cost-effective** – reduces data collection cost.
    - **Manageability** – smaller datasets are easier to handle.
    - **Representativeness** – if chosen properly, a sample can represent the population well and give reliable conclusions.

Types of Sampling Methods



A. Random Sampling Methods

1. Simple Random Sampling

- **Description**: Everyone has an equal chance of being selected.
- **Example**: Drawing names from a hat.
- **Pro**: Minimizes bias.
- **Con**: Can be impractical for large populations.

2. Stratified Sampling

- **Description**: Population is divided into groups, and samples are taken from each group.
- **Example**: Sampling different age groups separately.
- **Pro**: Ensures all groups are represented.
- **Con**: More complex to organize.

3. Systematic Sampling

- **Description**: Every nth person is selected.
- **Example**: Picking every 10th person on a list.
- **Pro**: Simple to implement.
- **Con**: Can be biased if there's a pattern in the list.

4. Cluster Sampling

- **Description**: Population is divided into clusters, and some clusters are fully sampled.
- **Example**: Sampling entire classrooms in a school.
- **Pro**: Cost-effective for large areas.
- **Con**: Higher chance of sampling error.

B. Non-Random Sampling Methods

1. Convenience Sampling

- **Description**: Choosing individuals who are easy to reach.

- **Example**: Surveying people in a mall.
- **Pro**: Quick and easy.
- **Con**: Likely to be biased.

2. **Judgmental (Purposive) Sampling**
   - **Description**: Choosing individuals based on specific criteria.
   - **Example**: Selecting experts in a field.
   - **Pro**: Useful for targeted studies.
   - **Con**: Can be biased by the researcher's judgment.
3. **Quota Sampling**
   - **Description**: Dividing the population into groups and taking a set number from each.
   - **Example**: Ensuring equal numbers of men and women in a study.
   - **Pro**: Ensures representation.
   - **Con**: Not random, can be biased.
4. **Snowball Sampling**
   - **Description**: Current participants recruit new participants.
   - **Example**: Existing survey participants asking friends to join.
   - **Pro**: Good for hard-to-reach groups.
   - **Con**: Can be biased and unrepresentative.

Errors in Hypothesis Testing

- When we perform hypothesis testing, we may sometimes make **wrong decisions** about the null hypothesis ($\mathbf{H_0}$).
- There are **two main types of errors**:
  - **Type I Error** → Rejecting a *true* null hypothesis (**false positive**)
  - **Type II Error** → Failing to reject a *false* null hypothesis (**false negative**)
- Another related idea is the **Standard Error of the Mean (SEM)**, which tells us how accurately a sample mean estimates the population mean.

Type I Error (False Positive)

Definition

- A Type I error occurs when the null hypothesis is true, but it is incorrectly rejected.
- Also known as a "false positive" or "alpha (α) error".

Characteristics

- **Significance Level (α)**: The probability of making a Type I error is denoted by **α**. Common choices are:
  - **α = 0.05** (5%)
  - **α = 0.01** (1%)
  - **α = 0.10** (10%)
- **Example**:
  - If **α = 0.05**, there is a **5% chance** of rejecting the null hypothesis when it is actually true.

Consequences

- Leads to believing an effect exists when it does not.
- Can result in incorrect scientific conclusions and wasted resources.

Pharmaceutical Example

- A company tests a new drug and concludes it is effective (**rejects H_0**) when in reality, it is **not effective**.
- This can result in **marketing an ineffective drug**, risking patient safety and increasing costs.

Type II Error (False Negative)

Definition

- A Type II error occurs when the null hypothesis is false, but it is incorrectly accepted (i.e., failing to reject H_0).
- Also known as a "false negative" or "beta (β) error".

Characteristics

1. **Probability (β)**: The probability of making a Type II error is denoted by **β**.
2. **Power of the Test**: Defined as **1 – β**, representing the probability of correctly rejecting a false null hypothesis.

Example:

- If **β = 0.20**, there is a **20% chance** of failing to detect a real effect.

Consequences

- Leads to **missing real effects** or **differences**.
- Can result in **discarding potentially valuable treatments** or interventions.

Pharmaceutical Example

- A company tests a new drug and concludes it is not effective (fails to reject H_0) when it actually is effective.
- This could cause a beneficial drug to be overlooked, delaying treatment for patients.

Standard Error of Mean (SEM)

Definition

- The Standard Error of the Mean (SEM) measures how precisely the sample mean (x̄) estimates the population mean (μ).
- In simple words:
  - it shows how much the **sample mean would vary from sample to sample**.

Formula

- When the population standard deviation (σ) is known:

$$\mathbf{SEM} = \frac{\boldsymbol{\sigma}}{\sqrt{\mathbf{n}}}$$

- When σ is unknown (use sample standard deviation s):

$$\mathbf{SEM} = \frac{\mathbf{s}}{\sqrt{\mathbf{n}}}$$

- Where:
  - **σ** = population standard deviation
  - **s** = sample standard deviation
  - **n** = sample size

Characteristics

- **Inversely related to sample size**
  - As n increases, SEM decreases.
  - This means the estimate of the mean becomes more precise.
- **Related to confidence intervals**
  - SEM is used to construct confidence intervals for the population mean.
  - It helps give a range in which the true mean is likely to lie.

SEM – graphical idea (sampling distribution of the mean)

- The curve shows the distribution of **sample means**.
- The **spread of this curve** is measured by the **SEM**.
- A smaller spread means a **smaller SEM** and better precision.

Importance in Hypothesis Testing

- The SEM plays a crucial role in hypothesis testing by:
  1. Helping to determine the test statistic (e.g., t-value, z-value) for comparing sample means.

2. Influencing the width of confidence intervals, which affects the decision to reject or fail to reject the null hypothesis.

Example in Pharmaceuticals

- A pharmaceutical company studies a new drug for reducing blood pressure.
- Given:
  - Sample size, **n = 100**
  - Sample standard deviation, **s = 10 mmHg**

Calculate SEM

$$\text{SEM} = \frac{s}{\sqrt{n}} = \frac{10}{\sqrt{100}} = \frac{10}{10} = 1 \text{ mmHg}$$

Interpretation

- The SEM is **1 mmHg**.
- This means:
  - the sample mean reduction in blood pressure is estimated with an average uncertainty of about **1 mmHg**.
- A **smaller SEM** would mean:
  - an even **more precise** estimate of the population mean.

- A pharmaceutical company tests a new drug to reduce blood pressure:
  - Sample size n = 100
  - Sample standard deviation s = 10 mmHg

Parametric test

Parametric Tests in Biostatistics

Introduction

- **Parametric tests** are statistical tests that:
  - assume the data follow a **specific distribution** (usually the **normal distribution**), and
  - use population parameters such as the **mean** and **variance**.
- These tests are very common in **biostatistics**, especially when the required assumptions are satisfied

Key Assumptions of Parametric Tests

1. **Normality**: Data should be normally distributed.
2. **Interval or Ratio Data**: Data must be on a scale that allows for meaningful calculation of mean and variance.
3. **Homogeneity of Variance (Homoscedasticity)**: For multi-group comparisons (e.g., ANOVA), the variances across groups should be approximately equal.

Advantages

- **Efficiency**: More powerful when assumptions are met.
- **Broad Application**: Widely used in statistical modeling, including regression and clinical trials.

Disadvantages

- **Assumption Sensitivity**: Violating assumptions may invalidate results.
- **Limited with Non-Normal Data**: Less suitable for skewed or ordinal data.

Common Parametric Tests

1. T-Tests

- Compare means between groups (e.g., one-sample, independent two-sample, and paired).

2. ANOVA (Analysis of Variance)

- Compares means across **three or more groups**.

3. Regression Analysis

- Models the relationship between a **dependent variable** and one or more **independent variables**.

T-Test: Overview and Types

- A **t-test** is used to check whether there is a statistically significant difference between means.
- It assumes:
  - the data come from a normal population, and
  - the test statistic follows a **t-distribution** when the population variance is unknown.

Types of t-Tests

1. One-Sample T-Test

Purpose

- Compare the **sample mean** to a known or hypothesized **population mean**.

Formula

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

- Where:
  - $\overline{x}$: Sample mean
  - μ: Population mean
  - s: Sample standard deviation
  - n: Sample size

Application Example

- A biotech company tests whether the **mean blood pressure** of patients on a new drug differs from **120 mmHg**.

2. Independent Two-Sample T-Test

Purpose

- Compare the means of two independent groups.

Types

- **Equal Variances (Pooled t-test)**
  - Assumes both groups have the same variance.
- **Unequal Variances (Welch's t-test)**
  - No assumption of equal variance.

Pooled T-Test Formula

$$t = \frac{\overline{x}_1 - \overline{x}_2}{s_p \cdot \sqrt{\frac{2}{n}}}$$

- Where:
  - $\overline{x}_1, \overline{x}_2$: Sample means
  - s_p: Pooled standard deviation
  - n: Sample size per group

Welch's T-Test Formula

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- Where:
  - s_1^2, s_2^2 : Sample variances
  - n1, n2 : Sample sizes

Application Example

- Compare mean serum cholesterol levels between men and women.

3. Paired T-Test

Purpose

- Compare means from two related samples, such as measurements before and after treatment.

Formula

$$t = \frac{\overline{d}}{s_d/\sqrt{n}}$$

- Where:
  - $\overline{d}$: Mean of the differences
  - s_d: Standard deviation of the differences
  - n: Number of pairs

Application Example

- Compare blood pressure before and after medication for the same group of patients.

Decision Rule in All T-Tests

- Compare the calculated t-value to the critical value from the t-distribution table.
- If t ≥ critical value → Reject H_0
- Indicates a statistically significant difference.

ANOVA (Analysis of Variance)

- **ANOVA** is a statistical method used to check whether the **means of three or more independent groups** are different.
- It works by comparing:
  - the variation **between groups**, and
  - the variation **within groups**.
- If the variation between groups is much larger than within groups, the group means are likely different.

One-Way ANOVA

Purpose

- To test the effect of one categorical factor on one continuous outcome.

Assumptions

- Observations are independent and randomly selected.
- Data in each group are normally distributed.
- Equal variances across groups (homogeneity of variance).

Application example

- A researcher studies whether three different diets produce different amounts of weight loss.
  - **Factor** → type of diet
  - **Response variable** → weight loss

ANOVA components

- The total variability in the data is divided into:
- **Between-group variability (SSB)**
  - due to differences between group means.
- **Within-group variability (SSW)**
  - due to variation inside each group.

F-ratio

$$F = \frac{MSB}{MSW}$$

- Where:
  - $MSB = \frac{SSB}{df_{between}}$ → mean square between groups
  - $MSW = \frac{SSW}{df_{within}}$ → mean square within groups
  - **SSB** = sum of squares between

- **SSW** = sum of squares within
- **df** = degrees of freedom

Interpretation

- A **large F-value** means:
  - group means are more different than would be expected by chance.

Two-Way ANOVA

Purpose

- **To study the effect of:**
  - two categorical factors, and
  - their interaction
- **on a continuous outcome.**

Assumptions

- Observations are **independent**.
- Residuals are **normally distributed**.
- Variances are **equal** across all groups.
- A **factorial design** is used (all combinations of factor levels are included).

Application example

- A biologist studies how:
  - **temperature** and
  - **humidity**
    affect **plant growth**.
  - **Factor A** → temperature
  - **Factor B** → humidity
  - **Response variable** → growth rate

ANOVA components (two-way)

- The total variation is divided into:
  - variation due to **Factor A**
  - variation due to **Factor B**
  - variation due to the **interaction (A × B)**
  - **error (residual)** variation

F-ratios

$$F_A = \frac{MS_A}{MS_{error}}, F_B = \frac{MS_B}{MS_{error}}, F_{AB} = \frac{MS_{AB}}{MS_{error}}$$

- Where:
  - MS_A= mean square for Factor A
  - MS_B= mean square for Factor B
  - MS_{AB}= mean square for interaction
  - MS_{error}= mean square of error

What does interaction mean? (simple words)

- **Interaction means:**
  - the effect of one factor depends on the level of the other factor.

Why use ANOVA?

- It allows testing many group means at the same time.
- It avoids doing many separate t-tests.
- It tells us:
  - whether at least one group mean is different.
- In **two-way ANOVA**, it also tells us:
  - the individual effects of both factors, and
  - whether an **interaction effect** exists.

Least Significant Difference (LSD) Test

Definition

- The Least Significant Difference (LSD) test is a post-hoc test used after ANOVA.
- It is applied only when the ANOVA F-test is significant.
- Its main job is to find which specific group means are different from each other.

Purpose

- Conducted only after a significant ANOVA result (i.e., a significant F-ratio).
- Used to make pairwise comparisons between group means.
- Especially helpful when researchers have specific hypotheses about differences between particular groups.

When to Use the LSD Test

- After performing a **one-way ANOVA**

- When ANOVA indicates at least **one group mean is different**
- To find **which groups differ** from each other
- When sample sizes are equal or approximately equal

LSD Formula

$$LSD = t_{\text{critical}} \times \sqrt{2 \times MSE \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

- Where:
  - t_{critical} = critical value from the *t-distribution* (based on ANOVA degrees of freedom and chosen confidence level)
  - **MSE** = Mean Square Error from the ANOVA table
  - n_1, n_2 = sample sizes of the two groups being compared

How It Works

1. **Calculate LSD value** using the formula.
2. **Compute the difference** between each pair of group means.
3. **Compare** the absolute difference to the LSD:
   - $If\ |\overline{x_1} - \overline{x_2}| >$ LSD
     - → The difference is **statistically significant**.
   - $|\overline{x_1} - \overline{x_2}| \leq$ LSD
     - → The difference is **not significant**.

Assumptions

1. **Independence**: Observations in each group must be independent.
2. **Normality**: Data in each group should be normally distributed.
3. **Homogeneity of Variance**: Variances across groups should be approximately equal.

Application Example

- After a significant one-way ANOVA comparing **three diets** on weight loss, the researcher uses the LSD test to compare:
  - Diet A vs Diet B
  - Diet A vs Diet C
  - Diet B vs Diet C
- to find **which diets differ significantly**.

Advantages of the LSD Test

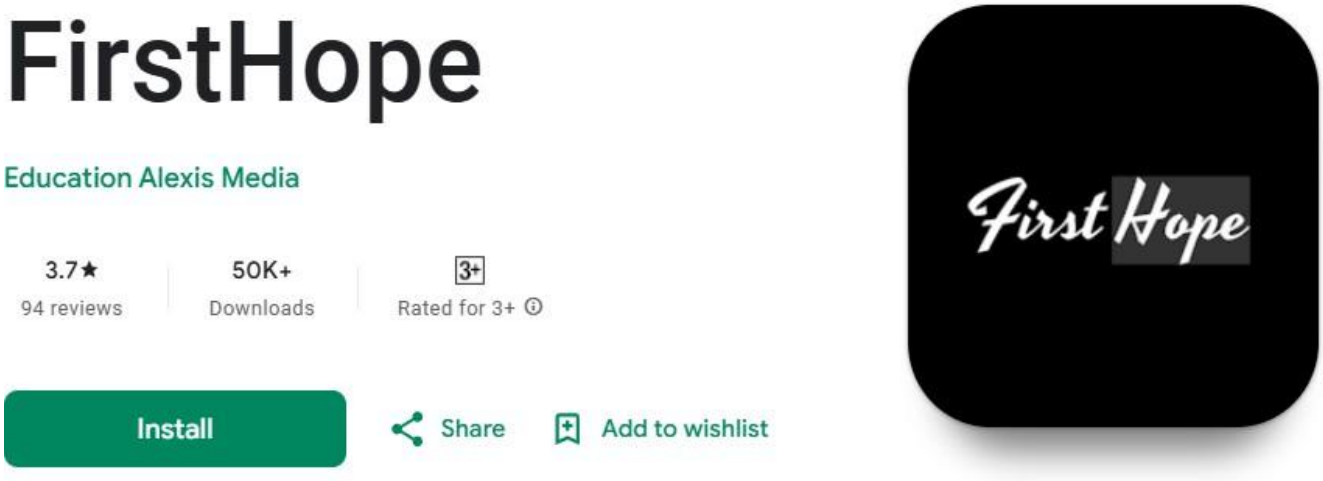
- Simple and easy to apply
- Easy to understand and interpret
- **More powerful** than some other post-hoc tests (such as Tukey's test) when:
    - only a few comparisons are made
- Useful when there are **specific pairwise hypotheses**

Limitations

- Has a **higher risk of Type I error** when many pairwise comparisons are made
- It is **less conservative** than methods such as:
    - Bonferroni correction
    - Tukey's HSD

First Hope

Easy Short Notes Designed to Help You Score Better



Android Users:

👉 Download on Google Play

iOS Users (Org Code- DYFFUX):

👉 Download on App Store

Connect With FirstHope

Visit Our Official Website ↗

Watch Free Videos on YouTube ↗

Join Our WhatsApp Community ↗

Join Our Telegram Channel ↗