

Statistics

Statistics

- The science of collecting, analyzing, presenting, and interpreting data.
- Data are the facts and figures that are collected, analyzed, and summarized for presentation and interpretation.
- Data may be classified as either quantitative or qualitative.
- **Quantitative data** measure either how much or how many of something, and **qualitative data** provide labels, or names, for categories of like items.

Statistics

- **Sample survey methods** are used to collect data from observational studies, and experimental design methods are used to collect data from experimental studies.
- The area of **descriptive statistics** is concerned primarily with methods of presenting and interpreting data using **graphs, tables, and numerical summaries**.
- Whenever statisticians use data from a sample—i.e., a subset of the population—to make statements about a population, they are performing **statistical inference**.
- **Estimation and hypothesis testing** are procedures used to make statistical inferences.

Statistical inference

- Statistical inference is the process of using data analysis to make conclusions about a population based on a sample.
- It involves hypothesis testing, confidence intervals, and probability-based decision-making.
- Real-life applications include medical research, business analytics, and social sciences.
- For ex:

A pharmaceutical company has developed a new vaccine for a seasonal flu virus. Before rolling it out to the entire population, they conduct a clinical trial on 10,000 volunteers.

Major approaches for statistical inference problem

Frequentist (classical) Inference:

- the unknown quantity θ is assumed to be a fixed quantity.
- θ is a deterministic (non-random) quantity that is to be estimated by the observed data.

$$\hat{\Theta} = \frac{Y}{n},$$

- n : population, Y : an event, $\hat{\Theta}$ is a random variable
- For example, in the polling problem we might consider θ as the percentage of people who will vote for a certain candidate, call him/her Candidate A. n is randomly chosen voters and Y is the number of people (among the randomly chosen voters) who say they will vote for Candidate A.

Major approaches for statistical inference problem

Bayesian Inference

- The unknown quantity Θ is assumed to be a random variable,
- Some initial guess about the distribution of Θ .
- After observing the data, we update the distribution of Θ using Bayes' Rule.
- For eg: consider the communication system in which the information is transmitted in the form of bits, i.e., 0's and 1's.
- Let's assume that, in each transmission, the transmitter sends a 1 with probability p , or it sends a 0 with probability $1-p$.

Random Sampling

The collection of random variables $X_1, X_2, X_3, \dots, X_n$ is said to be a **random sample** of size n if they are independent and identically distributed (i.i.d.), i.e.,

1. $X_1, X_2, X_3, \dots, X_n$ are independent random variables, and
2. they have the same distribution, i.e.,

$$F_{X_1}(x) = F_{X_2}(x) = \dots = F_{X_n}(x), \quad \text{for all } x \in \mathbb{R}.$$

Some Properties of Random Samples

- we assume that $X_1, X_2, X_3, \dots, X_n$ are a random sample. Specifically, we assume

1. the X_i 's are independent;
2. $F_{X_1}(x) = F_{X_2}(x) = \dots = F_{X_n}(x) = F_X(x)$;
3. $EX_i = EX = \mu < \infty$;
4. $0 < \text{Var}(X_i) = \text{Var}(X) = \sigma^2 < \infty$.

Sampling Distribution of the Sample Mean (\bar{X})

Let X_1, X_2, \dots, X_n be a random sample from a **normal population** with mean μ and variance σ^2 , i.e.,

$$X_i \sim N(\mu, \sigma^2)$$

1.1 Distribution of the Sample Mean

The **sample mean** is defined as:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

From probability theory, the sample mean follows:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

This means:

- The **mean** of \bar{X} is $E[\bar{X}] = \mu$, meaning the estimator is **unbiased**.
- The **variance** of \bar{X} is $Var(\bar{X}) = \frac{\sigma^2}{n}$, meaning as sample size n increases, variance decreases.
- If the original population is normally distributed, the **sample mean is also normally distributed** for any n .

Sample Mean

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}.$$

Properties of the sample mean

1. $E\bar{X} = \mu$.
2. $\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$.
3. Weak Law of Large Numbers (WLLN):

$$\lim_{n \rightarrow \infty} P(|\bar{X} - \mu| \geq \epsilon) = 0.$$

4. Central Limit Theorem: The random variable

$$Z_n = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{X_1 + X_2 + \dots + X_n - n\mu}{\sqrt{n}\sigma}$$

converges in distribution to the standard normal random variable as n goes to infinity, that is

$$\lim_{n \rightarrow \infty} P(Z_n \leq x) = \Phi(x), \quad \text{for all } x \in \mathbb{R}$$

where $\Phi(x)$ is the standard normal CDF.

- Suppose a population is normally distributed with mean $\mu=50$ and variance $\sigma^2=25$. A sample of size $n=10$ is taken. Find:
 - 1.The distribution of the sample mean \bar{X} .
 - 2.The probability that $\bar{X} > 52$.
 - 3.The distribution of the sample variance s^2

Z-Table (Standard Normal Table)

- The table gives $P(Z < z)$, which is the cumulative probability from $-\infty$ to z .
- If you need $P(Z > z)$, use:
 - $P(Z > z) = 1 - P(Z < z)$
- If you need **$P(a < Z < b)$** , use:
 - $P(a < Z < b) = P(Z < b) - P(Z < a)$

A small portion of the Z-table

Z-score	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177

- For **Z = 1.26**, find the row for **1.2** and the column for **0.06**, which gives **0.8962**.

- Thus,

$$P(Z < 1.26) = 0.8962$$

$$P(Z > 1.26) = 1 - 0.8962 = 0.1038$$

Example

- A normal population has $\mu=80$ and $\sigma=12$.
- Find $P(75 < \bar{X} < 85)$ for a sample size of $n=16$.

Order Statistics

- Given a random sample, we might be interested in quantities such as the largest, the smallest, or the middle value in the sample.
- let $X_1, X_2, X_3, \dots, X_n$ be a random sample from a continuous distribution with CDF $F_X(x)$.
- Let us order X_i 's from the smallest to the largest and denote the resulting sequence of random variables as

$$X_{(1)}, X_{(2)}, \dots, X_{(n)}.$$

Thus, we have

$$X_{(1)} = \min (X_1, X_2, \dots, X_n);$$

and

$$X_{(n)} = \max (X_1, X_2, \dots, X_n).$$

We call $X_{(1)}, X_{(2)}, \dots, X_{(n)}$ the **order statistics** of the random sample $X_1, X_2, X_3, \dots, X_n$.

Theorem 8.1

Let X_1, X_2, \dots, X_n be a random sample from a continuous distribution with CDF $F_X(x)$ and PDF $f_X(x)$. Let $X_{(1)}, X_{(2)}, \dots, X_{(n)}$ be the order statistics of $X_1, X_2, X_3, \dots, X_n$. Then the CDF and PDF of $X_{(i)}$ are given by

$$f_{X_{(i)}}(x) = \frac{n!}{(i-1)!(n-i)!} f_X(x) [F_X(x)]^{i-1} [1 - F_X(x)]^{n-i},$$

$$F_{X_{(i)}}(x) = \sum_{k=i}^n \binom{n}{k} [F_X(x)]^k [1 - F_X(x)]^{n-k}.$$

Also, the joint PDF of $X_{(1)}, X_{(2)}, \dots, X_{(n)}$ is given by

$$f_{X_{(1)}, \dots, X_{(n)}}(x_1, x_2, \dots, x_n) = \begin{cases} n! f_X(x_1) f_X(x_2) \cdots f_X(x_n) & \text{for } x_1 \leq x_2 \leq x_2 \cdots \leq x_n \\ 0 & \text{otherwise} \end{cases}$$

Example

Let X_1, X_2, X_3, X_4 be a random sample from the $Uniform(0, 1)$ distribution, and let $X_{(1)}, X_{(2)}, X_{(3)}, X_{(4)}$. Find the PDFs of $X_{(1)}, X_{(2)}$, and $X_{(4)}$.

Here, the ranges of the random variables are $[0, 1]$, so the PDFs and CDFs are zero outside of $[0, 1]$. We have

$$f_X(x) = 1, \quad \text{for } x \in [0, 1],$$

and

$$F_X(x) = x, \quad \text{for } x \in [0, 1].$$

By [Theorem 8.1](#), we obtain

$$\begin{aligned} f_{X_{(1)}}(x) &= \frac{4!}{(1-1)!(4-1)!} f_X(x) [F_X(x)]^{1-1} [1 - F_X(x)]^{4-1} \\ &= 4f_X(x) [1 - F_X(x)]^3 \\ &= 4(1-x)^3, \quad \text{for } x \in [0, 1]. \end{aligned}$$

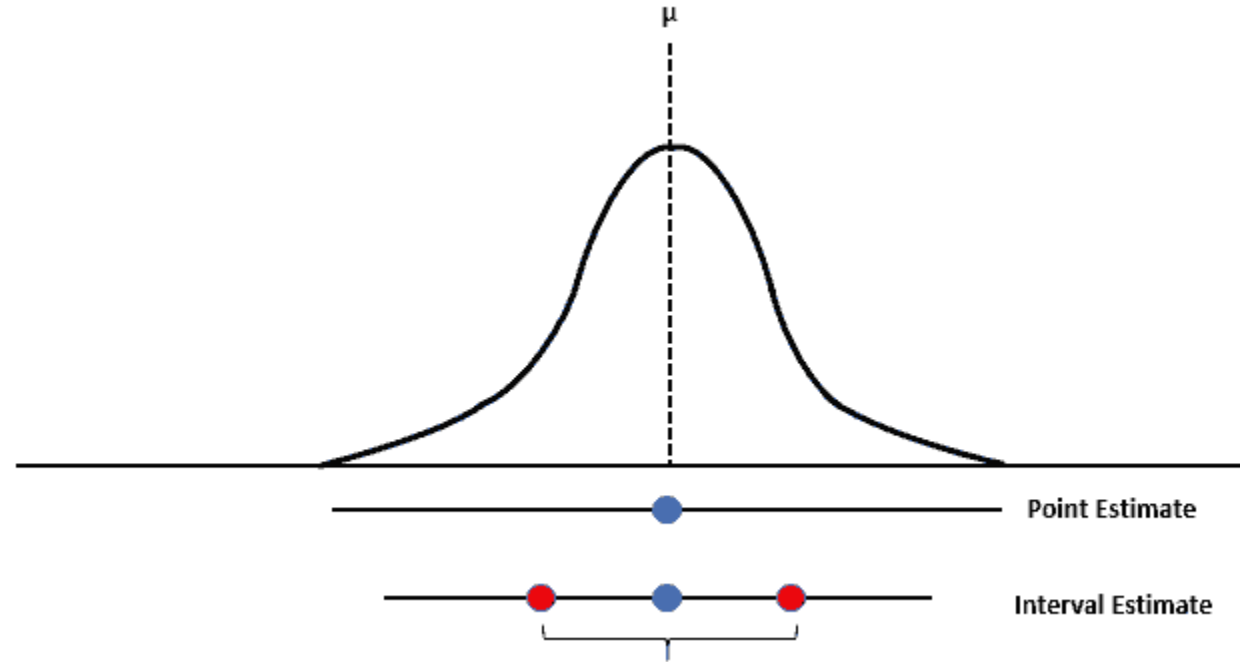
$$\begin{aligned} f_{X_{(2)}}(x) &= \frac{4!}{(2-1)!(4-2)!} f_X(x) [F_X(x)]^{2-1} [1 - F_X(x)]^{4-2} \\ &= 12f_X(x) F_X(x) [1 - F_X(x)]^2 \\ &= 12x(1-x)^2, \quad \text{for } x \in [0, 1]. \end{aligned}$$

$$\begin{aligned} f_{X_{(4)}}(x) &= \frac{4!}{(4-1)!(4-4)!} f_X(x) [F_X(x)]^{4-1} [1 - F_X(x)]^{4-4} \\ &= 4f_X(x) [F_X(x)]^3 \\ &= 4x^3, \quad \text{for } x \in [0, 1]. \end{aligned}$$

Estimation

- Estimation is a process to obtain the **values of unknown population parameters** with the help of sample data.
- Data analysis framework that combines effect sizes and confidence intervals to **plan an experiment, analyze data, and interpret the results.**
- Point and Interval estimates are the two forms of population parameter estimation based on sample data.

Point estimation & Interval estimation



Point estimation

- Point estimation provides a **single numerical value** as an estimate of an unknown population parameter. It does not indicate how much error might be involved in the estimation.
- For example, if we want to estimate the average income of all employees in a company, we can take a sample and calculate the **sample mean** as an estimate of the **population mean (μ)**.

Point Estimation

- Here, we assume that θ is an unknown parameter to be estimated.
- For example, θ might be the expected value of a random variable, $\theta = EX$. Θ is a fixed (non-random) quantity.
- To estimate θ , we define a point estimator $\hat{\Theta}$ that is a function of the random sample, i.e.,

$$\hat{\Theta} = h(X_1, X_2, \dots, X_n).$$

For example, if $\theta = EX$, we may choose $\hat{\Theta}$ to be the sample mean

$$\hat{\Theta} = \bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}.$$

Point Estimation

Mean (\bar{x}) → Estimates Population Mean (μ)

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i$$

Variance (s^2) → Estimates Population Variance (σ^2)

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

Standard Deviation (s) → Estimates Population Standard Deviation (σ)

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2}$$

Proportion (\hat{p}) → Estimates Population Proportion (p)

$$\hat{p} = \frac{x}{n}$$

Properties of Estimators

- Estimators should be **unbiased**.
 - expected value equals the true parameter value.
- The estimator should be **efficient**.
 - it has the **lowest variance** among all unbiased estimators of a parameter.
- An estimator should be **consistent**.
 - as the sample size increases, the estimated value gets closer to the true population parameter.
 - More data improves the accuracy of estimation.

Evaluating Estimators

- Three main desirable properties for point estimators

1. The **bias** of an estimator $\hat{\Theta}$ tells us on average how far $\hat{\Theta}$ is from the real value of θ .

Let $\hat{\Theta} = h(X_1, X_2, \dots, X_n)$ be a point estimator for θ . The **bias** of point estimator $\hat{\Theta}$ is defined by

$$B(\hat{\Theta}) = E[\hat{\Theta}] - \theta.$$

Let $\hat{\Theta} = h(X_1, X_2, \dots, X_n)$ be a point estimator for a parameter θ . We say that $\hat{\Theta}$ is an **unbiased** estimator of θ if

$$B(\hat{\Theta}) = 0, \quad \text{for all possible values of } \theta.$$

i. Unbiasedness

- An estimator $\hat{\theta}$ is **unbiased** if its expected value is equal to the true population parameter (θ):

$$E(\hat{\theta}) = \theta$$

- Example: The sample mean \bar{x} is an unbiased estimator of the population mean μ :

$$E(\bar{X}) = \mu$$

- Example of a biased estimator: The **sample variance (s^2)** is an **unbiased** estimator of **population variance (σ^2)** when we use **n-1** in the denominator, but **biased** when we divide by **n** instead of **n-1**.

iii. *consistency*

- In general, if $\hat{\Theta}$ is a point estimator for θ , we can write

$$\begin{aligned}MSE(\hat{\Theta}) &= E[(\hat{\Theta} - \theta)^2] \\&= \text{Var}(\hat{\Theta} - \theta) + (E[\hat{\Theta} - \theta])^2 \\&= \text{Var}(\hat{\Theta}) + B(\hat{\Theta})^2.\end{aligned}$$

If $\hat{\Theta}$ is a point estimator for θ ,

$$MSE(\hat{\Theta}) = \text{Var}(\hat{\Theta}) + B(\hat{\Theta})^2,$$

where $B(\hat{\Theta}) = E[\hat{\Theta}] - \theta$ is the bias of $\hat{\Theta}$.

iii. *consistency*

- An estimator $\hat{\theta}$ is **consistent** if it gets **closer to the true parameter (θ)** as the **sample size (n) increases**.

Let $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_n, \dots$, be a sequence of point estimators of θ . We say that $\hat{\theta}_n$ is a **consistent** estimator of θ , if

$$\lim_{n \rightarrow \infty} P(|\hat{\theta}_n - \theta| \geq \epsilon) = 0, \text{ for all } \epsilon > 0.$$

- Example: The **sample mean (\bar{x})** is a consistent estimator of μ because as we take larger samples, it converges to μ .

iii. Efficiency

- Among multiple unbiased estimators, the **most efficient** estimator has the **smallest variance**.

$$Var(\hat{\theta}_1) < Var(\hat{\theta}_2) \Rightarrow \hat{\theta}_1 \text{ is more efficient}$$

Example: If we have two estimators of μ , the one with the smaller variance is preferred.

Estimator 1: $Var(\hat{\theta}_1) = 5$

Estimator 2: $Var(\hat{\theta}_2) = 2$

Mean squared error (MSE)

The **mean squared error** (MSE) of a point estimator $\hat{\Theta}$, shown by $MSE(\hat{\Theta})$, is defined as

$$MSE(\hat{\Theta}) = E[(\hat{\Theta} - \theta)^2].$$

Example 8.3

Let $X_1, X_2, X_3, \dots, X_n$ be a random sample from a distribution with mean $EX_i = \theta$, and variance $\text{Var}(X_i) = \sigma^2$. Consider the following two estimators for θ :

1. $\hat{\Theta}_1 = X_1$.
2. $\hat{\Theta}_2 = \bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$.

Find $MSE(\hat{\Theta}_1)$ and $MSE(\hat{\Theta}_2)$ and show that for $n > 1$, we have

$$MSE(\hat{\Theta}_1) > MSE(\hat{\Theta}_2).$$

Examples

check whether the sample mean \bar{x} is an unbiased estimator of μ .

We take a sample of 5 people's salaries: (40k, 50k, 60k, 70k, 80k)

Compute \bar{x} :

ii. Checking Consistency

Consider two estimators for population mean μ :

- **Estimator A** uses a sample size of **10**.
- **Estimator B** uses a sample size of **100**.

iii. Comparing Efficiency

Two unbiased estimators for μ have the following variances:

- **Estimator 1:** $\text{Var}(\theta_1^{\wedge})=4$
- **Estimator 2:** $\text{Var}(\theta_2^{\wedge})=2$

Solution

$$\bar{X} = \frac{40 + 50 + 60 + 70 + 80}{5} = 60$$

- If we take repeated samples, the **expected value** of \bar{x} will approach the population mean μ , proving that it is unbiased.

ii. Checking consistency

- As **n increases**, the sample mean \bar{x} in **Estimator B** is **closer** to μ than in **Estimator A**, proving **consistency**.

iii. Since **Estimator 2 has a smaller variance**, it is more **efficient**.

Interval Estimation

- Interval estimation provides a **range of values** (instead of a single value) that is likely to contain the unknown population parameter.
- This range is called a **confidence interval (CI)** and is associated with a confidence level (e.g., 95%).
- For example, instead of estimating the population mean as $\bar{x} = 50000$, we say,
"The true population mean is between 48000 and 52000 with 95% confidence."

Finding Interval Estimators

- Confidence level = $1-\alpha$
- Let X be a continuous random variable with CDF $F_X(x)=P(X\leq x)$. Suppose that we are interested in finding two values x_h and x_l such that

$$P\left(x_l \leq X \leq x_h\right) = 1 - \alpha.$$

One way to do this, is to choose x_l and x_h such that

$$P(X \leq x_l) = \frac{\alpha}{2}, \quad \text{and} \quad P(X \geq x_h) = \frac{\alpha}{2}.$$

Equivalently,

$$F_X(x_l) = \frac{\alpha}{2}, \quad \text{and} \quad F_X(x_h) = 1 - \frac{\alpha}{2}.$$

We can rewrite these equations by using the inverse function F_X^{-1} as

$$x_l = F_X^{-1}\left(\frac{\alpha}{2}\right), \quad \text{and} \quad x_h = F_X^{-1}\left(1 - \frac{\alpha}{2}\right).$$

We call the interval $[x_l, x_h]$ a $(1 - \alpha)$ interval for X . Figure 8.2 shows the values of x_l and x_h using the CDF of X , and also using the PDF of X .

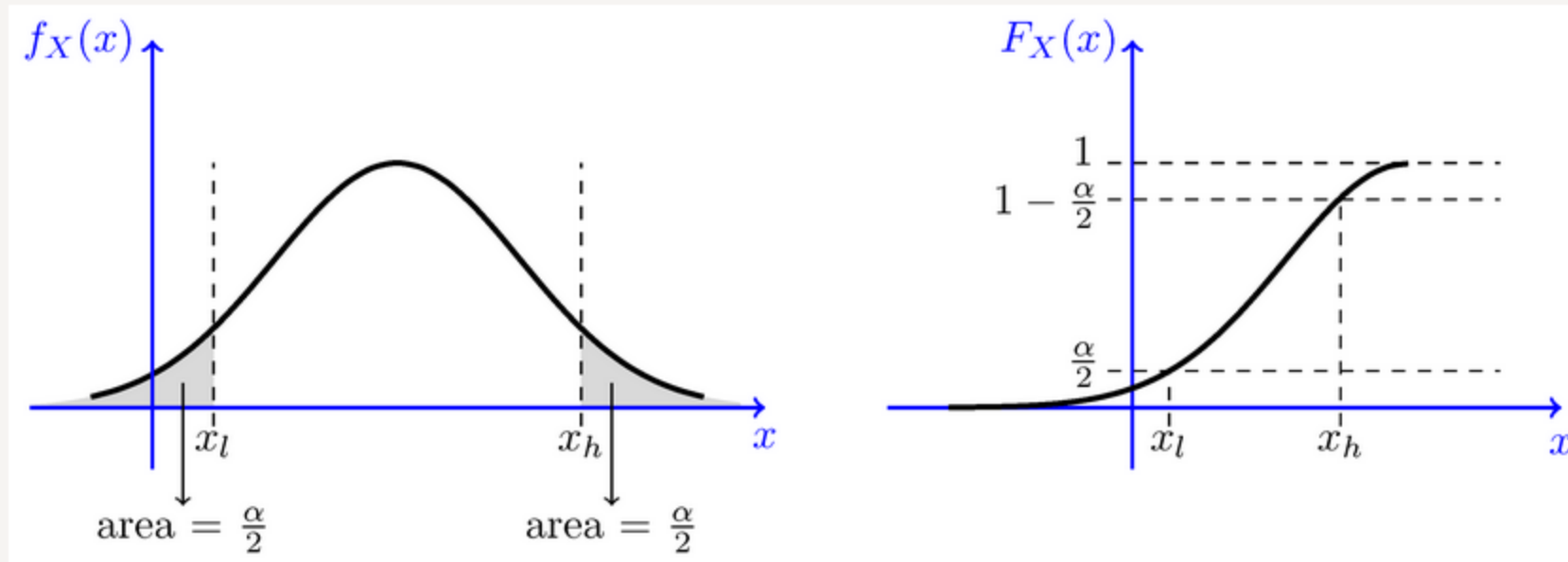


Figure 8.2 - $[x_l, x_h]$ is a $(1 - \alpha)$ interval for X , that is, $P(x_l \leq X \leq x_h) = 1 - \alpha$.

Example

Let $Z \sim N(0, 1)$, find x_l and x_h such that

$$P\left(x_l \leq Z \leq x_h\right) = 0.95$$

Confidence interval

Now suppose that $X_1, X_2, X_3, \dots, X_n$ is a random sample from a distribution with *unknown* variance $\text{Var}(X_i) = \sigma^2$. Our goal is to find a $1 - \alpha$ confidence interval for $\theta = EX_i$. We also assume that n is large. By the above discussion, we can say

$$P\left(\bar{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \theta \leq \bar{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

Sampling Distributions: Chi-Square, t, F, and Z

- **Confidence Intervals for Normal Samples when n is small**
- In statistics, **sampling distributions** describe the probability distributions of sample statistics (such as means or variances) when samples are drawn from a population.
- The some more common sampling distributions are:
 - Z-distribution (Standard Normal Distribution)
 - t-distribution (Student's t)
 - Chi-square (χ^2) distribution
 - F-distribution

Gamma Distribution

- The **Gamma distribution** is a continuous probability distribution used to model waiting times in queuing systems, life spans of objects, and other real-world processes.
- A random variable X follows a Gamma distribution with shape parameter k and scale parameter θ , denoted as:

$$X \sim \text{Gamma}(k, \theta)$$

The probability density function (PDF) is:

$$f_X(x) = \frac{x^{k-1} e^{-x/\theta}}{\theta^k \Gamma(k)}, \quad x > 0$$

where $\Gamma(k)$ is the **Gamma function**, defined as:

$$\Gamma(k) = \int_0^{\infty} t^{k-1} e^{-t} dt$$

For integer values of k , the Gamma function simplifies to:

$$\Gamma(k) = (k - 1)!$$

Mean (Expected Value):

$$E[X] = k\theta$$

Variance:

$$\text{Var}(X) = k\theta^2$$

Chi-Square (χ^2) Distribution

- A continuous probability distribution commonly used in hypothesis testing, confidence intervals, and statistical inference.
- A random variable X follows a **Chi-Square distribution** with k degrees of freedom (df), denoted as:

$$X \sim \chi^2(k)$$

The probability density function (PDF) is:

$$f_X(x) = \frac{x^{(k/2)-1} e^{-x/2}}{2^{k/2} \Gamma(k/2)}, \quad x > 0$$

where $\Gamma(k/2)$ is the **Gamma function**.

Properties (χ^2) Distribution

Mean (Expected Value):

$$E[X] = k$$

Variance:

$$\text{Var}(X) = 2k$$

Skewness:

$$\text{Skewness} = \frac{2}{\sqrt{k}}$$

Special Cases:

- If $k = 1$, the Chi-Square distribution is **highly skewed**.
- As k increases, the distribution approaches a **normal distribution** (by the **Central Limit Theorem**).

Example

Let $X \sim \chi^2(10)$. Find:

1. $E[X]$

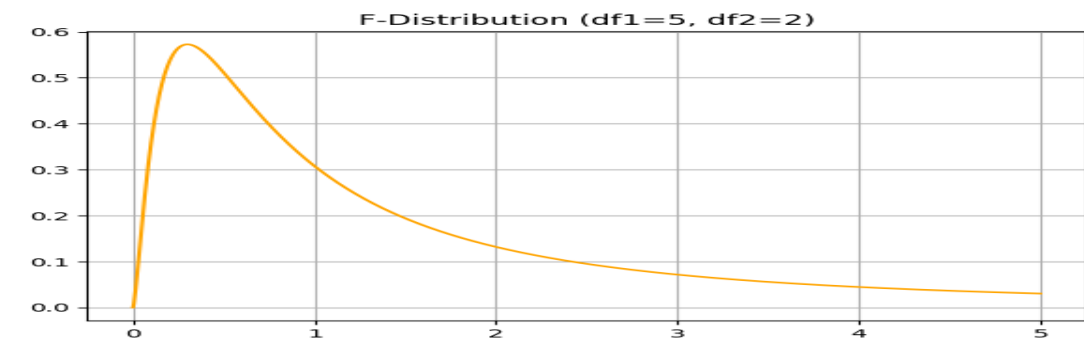
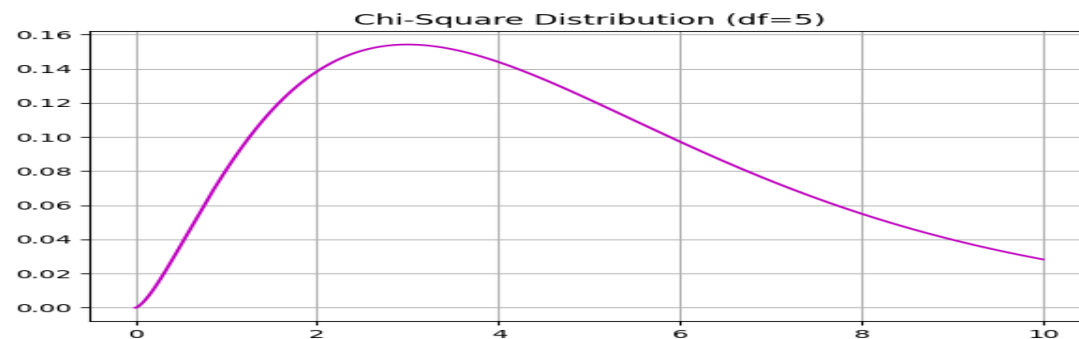
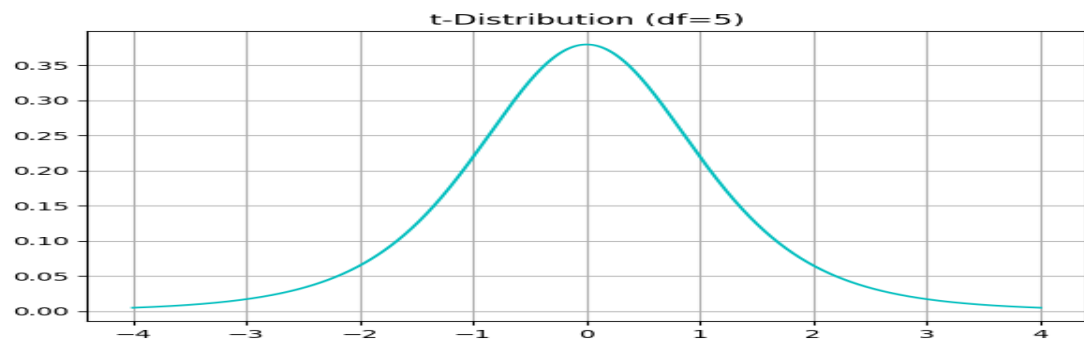
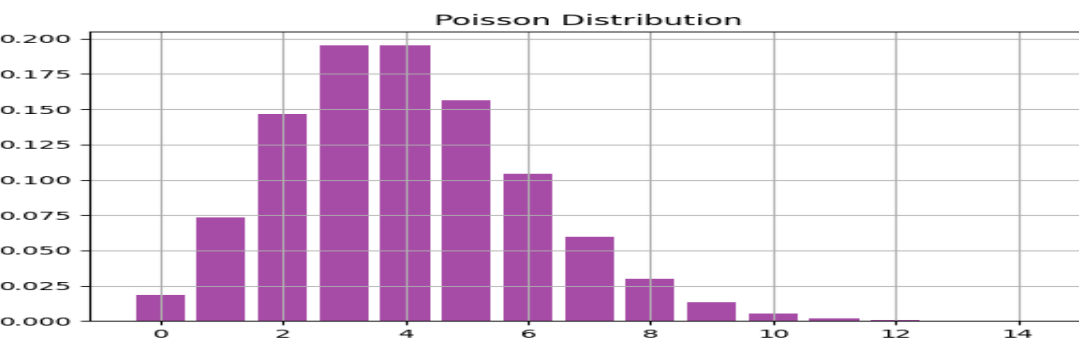
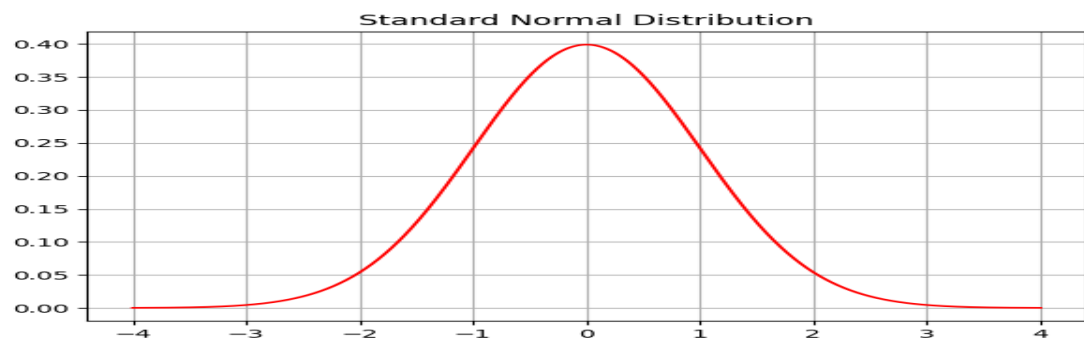
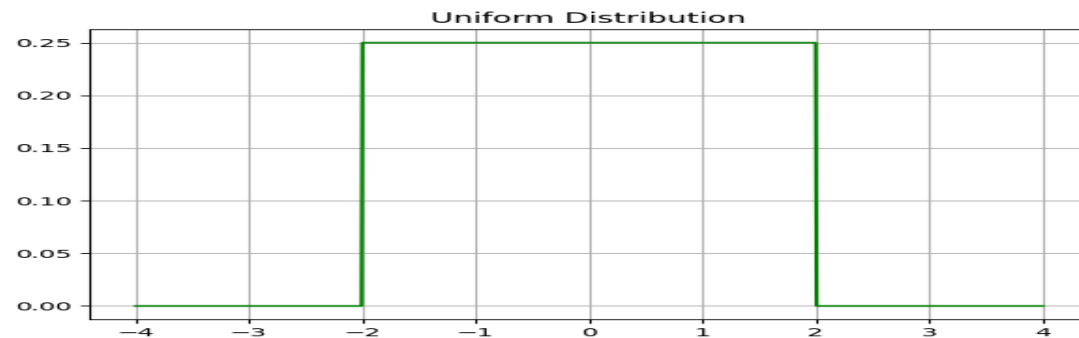
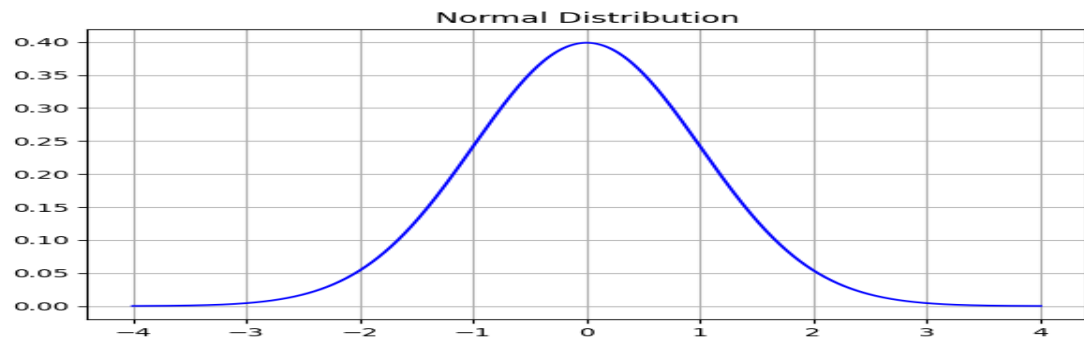
2. $\text{Var}(X)$

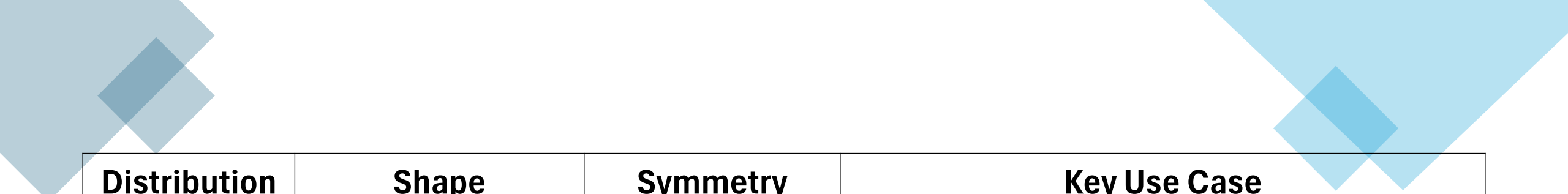
Solution:

1. $E[X] = k = 10$


2. $\text{Var}(X) = 2k = 20$

Common Probability Distributions





Distribution	Shape	Symmetry	Key Use Case
Normal	Bell-shaped	Symmetric	General-purpose modeling, errors, natural phenomena
Uniform	Flat	Symmetric	Equal probability scenarios, randomness
Standard Normal	Bell-shaped	Symmetric	Z-score analysis, normalization
Poisson	Skewed (right)	Asymmetric	Count-based data, rare events
t	Bell-shaped with heavier tails	Symmetric	Small-sample hypothesis testing
Chi-Square	Skewed (right)	Asymmetric	Variance testing, categorical data analysis
F	Skewed (right)	Asymmetric	Comparing variances, ANOVA



Z-Distribution (Standard Normal Distribution)

- The population standard deviation (σ) is known, or the sample size is large ($n \geq 30$).
- Bell-shaped and symmetric.
- Distribution $N(0,1)$
- Used for large sample hypothesis testing and confidence intervals.
- Eg: If the heights of students follow a normal distribution with $\mu = 170$ cm and $\sigma = 10$ cm, and we take a sample of $n = 50$, we can use the Z-distribution to find the probability of a sample mean greater than 172 cm.

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

Z-test

- **Problem Statement**
- A company claims that the average weight of a packaged product is **500 grams**. A random sample of **30 packages** is taken, and the sample has a **mean weight of 495 grams** with a **standard deviation of 10 grams**. At a **5% significance level ($\alpha=0.05$)**, test whether the company's claim is valid using a **one-sample Z-test**.

Step 1: Define Hypotheses

- Step 1: Define Hypotheses
 - **Null Hypothesis (H_0):** The mean weight of the product is **500 grams**.
 - $H_0: \mu = 500$
 - **Alternative Hypothesis (H_A):** The mean weight is **not 500 grams** (two-tailed test). $H_A: \mu \neq 500$
- **Step 2: Identify Given Data**
 - Population mean (μ_0) = **500 grams**
 - Sample mean (\bar{X}) = **495 grams**
 - Sample standard deviation (s) = **10 grams**
 - Sample size (n) = **30**
 - Significance level (α) = **0.05**

- Step 3: Calculate the Z-Statistic
- **Step 4: Find the Critical Value**
 - For a **two-tailed test** at **$\alpha=0.05$** , the critical value from the Z-table
- Step 5: Decision Rule

Chi-Square Test (χ^2 -test)

- It is used to analyze **categorical data**.
- It determines whether observed frequencies differ significantly from expected frequencies.
- Types of Chi-Square Tests
 - Chi-Square Goodness-of-Fit Test
 - Tests if a sample follows a specific distribution.
 - Chi-Square Test for Independence
 - Tests if two categorical variables are related.

Chi-Square Goodness-of-Fit Test

- Used when you have one categorical variable and want to compare observed data to expected data.

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

- where:
- O_i = Observed frequency
- E_i = Expected frequency

Example

- A factory produces **4 different types of eco-friendly packaging** (A, B, C, and D). The company claims that the packaging types are equally preferred by customers. A sample of **200 customers** is surveyed, and the observed preferences are recorded as follows:

Packaging Type	Observed Frequency (O)
A	55
B	45
C	50
D	50
Total	200

- We want to test whether customer preferences follow the assumed **uniform distribution** at a **5% significance level ($\alpha = 0.05$)**.

Solution

- **Step 1: State the Hypotheses**
 - **Null Hypothesis (H0):** Customer preferences are uniformly distributed across packaging types.
 - **Alternative Hypothesis (H1):** Customer preferences are not uniformly distributed.
- **Step 2:** Calculate the expected frequency for each product
- **Step 3:** Compute the Chi-Square Statistic
- **Step 4:** Determine the Critical Value
- **Step 5:** Make a Decision

Chi-Square Test for Independence

- Used when analyzing **two categorical variables** in a contingency table.

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

O_{ij} = Observed frequency in row i , column j

E_{ij} = Expected frequency, calculated as:

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

Example

A company surveys 200 customers about their preferred payment method by gender. We want to test whether payment method preference depends on gender using a Chi-Square test for independence at a 5% significance level.

Payment Method	Male	Female	Total
Credit Card	40	50	90
UPI	35	25	60
Cash	25	25	50
Total	100	100	200

Solution

- Step 1: State Hypotheses
 - Null Hypothesis (H_0): Payment method preference is independent of gender.
 - Alternative Hypothesis (H_1): Payment method preference is dependent on gender.
- Step 2: Compute Expected Frequencies
- Step 3: Compute Chi-Square Statistic
- Step 4: Determine the Critical Value
- **Step 5: Make a Decision**
 - If $\chi^2_{\text{calculated}} < \text{critical value}$, fail to reject H_0
 - If $\chi^2_{\text{calculated}} \geq \text{critical value}$, reject H_0 (association exists).

t-Distribution (Student's t)

- The **population standard deviation** (σ) is unknown, and the **sample size is small** ($n < 30$).
- Similar to the normal distribution but wider (heavier tails).
- The shape depends on degrees of freedom ($df = n - 1$).
- As n increases, it approaches the Z-distribution
- **Example:** If we take a small sample ($n = 10$) from a population where σ is unknown, we use the **t-distribution** instead of the Z-distribution.

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

$$s = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n - 1}}$$

Example

- A company wants to check if the average battery life of its product is different from 10 hours. A random sample of 15 batteries is tested, and the sample mean is 9.5 hours with a sample standard deviation of 1.2 hours. Can we conclude that the true mean is different from 10 hours at a 5% significance level?

Solution

- **Null Hypothesis (H0):** $\mu=10$ (no difference in mean).
- **Alternative Hypothesis (H1):** $\mu \neq 10$ (battery life is different).
- **Compute the t-statistic:**
 - $t=-1.61$
- **Find the Critical Value**
 - From the t-table, for $df=14$ at $\alpha=0.05$, $t_{critical}=\pm 2.145$.
- **Decision:** Since $-2.145 < -1.61 < 2.145$, we fail to reject H0.
 - There is **not enough evidence** to conclude that the battery life is different from 10 hours.

Hypothesis Testing

- Hypothesis testing is a statistical method used to make inferences about a population based on sample data. It involves:
 - Null Hypothesis (H_0): Assumes no effect or difference.
 - Alternative Hypothesis (H_a): Assumes there is an effect or difference.
 - Test Statistic: A standardized value (e.g., Z-score, t-score).
 - Significance Level (α): Common values are 0.05 (5%) or 0.01 (1%).
 - Decision Rule:
 - If the p-value $< \alpha$, reject H_0 .

Type I and Type II Errors in Hypothesis Testing

Type I and Type II Error		
Null hypothesis is ...	True	False
Rejected	Type I error False positive Probability = α	Correct decision True positive Probability = $1 - \beta$
Not rejected	Correct decision True negative Probability = $1 - \alpha$	Type II error False negative Probability = β

Type I and Type II Errors in Hypothesis Testing

- Type 1 and type 2 errors are mistakes that can occur when testing a hypothesis. They are also known as alpha and beta errors.
- Type 1 error
 - Also known as a false positive
 - Occurs when a null hypothesis is rejected even though it is true
 - For example, finding a suspect guilty when they are innocent
- Type 2 error
 - Also known as a false negative
 - Occurs when a null hypothesis is not rejected even though it is false
 - For example, finding a suspect innocent when they are guilty

Example Scenarios

- **Scenario 1: Medical Diagnosis**
 - **H0:** A patient does not have a disease.
 - **H_a:** The patient has the disease.
 - **Type I Error:** Diagnosing a healthy patient as sick (False Positive).
 - **Type II Error:** Failing to detect a disease in a sick patient (False Negative).
- **Scenario 2: Quality Control in Manufacturing**
 - **H0:** A machine produces defect-free products.
 - **H_a:** The machine produces defective products.
 - **Type I Error:** Stopping the machine when it is actually working fine.
 - **Type II Error:** Letting defective products pass the inspection.

When to use different significance levels

- If the consequences of a type 1 error are more severe, you can lower the significance level to reduce the chance of rejecting a true null hypothesis.
- If the consequences of a type 2 error are more severe, you can increase the significance level to improve the statistical power of your tests.

One-Tailed vs. Two-Tailed Tests

One-Tailed Test: Tests if a parameter is **greater than or less than** a certain value.

- Example: Testing if a new drug increases survival rate.

Two-Tailed Test: Tests if a parameter is **different** from a certain value in **either direction**.

- Example: Checking if a new teaching method affects students' test scores (could be better or worse).

One-Tailed Test

Types of One-Tailed Tests

- **Right-Tailed Test (Upper-Tailed Test):** Used when we want to test if a **parameter is greater than the hypothesized value**.
 - Example: A company claims that a new machine increases production speed beyond 50 units/hour.
- **Left-Tailed Test (Lower-Tailed Test):** Used when we want to test if a **parameter is less than the hypothesized value**.
 - Example: A manufacturer wants to verify if a battery's average life is lower than 500 hours.

Two-Tailed Test

- A **two-tailed test** checks whether a sample statistic is significantly **different from** the hypothesized value in **either direction** (greater or lesser). It is used when we have a non-directional hypothesis.
- **Example of a Two-Tailed Test**
- A company wants to check if the average weight of a product is **exactly** 500g or **significantly different** (could be higher or lower).
- **Hypotheses:**
 - **H₀:** $\mu=500$ (Product weight is 500g)
 - **H_a:** $\mu \neq 500$ (Product weight is different from 500g)
- **Rejection Region:** If the test statistic falls in **either** tail of the distribution.

Analysis of Variance (ANOVA)

Analysis of Variance (ANOVA)

- ANOVA is used to compare the means of **two or more groups** to determine if there is a significant difference between them.

One-Way ANOVA

- Used when comparing **one independent variable** (factor) with **multiple groups**.
- Example: Comparing the average test scores of students from **three different schools**.

Two-Way ANOVA

- Used when analyzing the effect of **two independent variables** on a dependent variable.
- Example: Checking how **study method (factor 1)** and **class size (factor 2)** affect students' exam performance.

How ANOVA Works: The F-Test

- ANOVA calculates the **F-statistic**, which is the ratio of:

$$F = \frac{\text{Variation between groups}}{\text{Variation within groups}}$$

- A higher F-value suggests that group means are more different than expected by random chance.
- We compare F to a critical value from the F-distribution to determine significance.

Assumptions of ANOVA

- Independence: Observations should be independent.
- Normality: Data in each group should be normally distributed.
- Homogeneity of Variance (Homoskedasticity): Variances of different groups should be similar.

One-Way ANOVA Calculation

Problem Statement

A researcher tests the effect of three teaching methods on student performance. The test scores of students are:

Teaching Method	Scores
Method A	78, 85, 88, 92, 75
Method B	82, 79, 84, 86, 88
Method C	91, 89, 94, 96, 85

Conduct a **One-Way ANOVA** to check if there is a significant difference between the means of the three methods.

Solution

- **Step 1: Compute the Mean for Each Group**
 - Mean of **Method A**: \bar{X}_A
 - Mean of **Method B**: \bar{X}_C
 - Mean of **Method C**: \bar{X}_B
- Step 2: Compute the Total Mean (\bar{X}_{total})
- **Step 3: Compute the Sum of Squares (SS): Total sum of square=SSB+SSW**
 - **Between-Groups Variability (SSB)**: Measures variation **between** the group means.

$$SSB = n \sum (\bar{X}_i - \bar{X}_{\text{overall}})^2$$

- **Within-Groups Variability (SSW)**: Measures variation **within** each group.
- $$SSW = \sum (X_{ij} - \bar{X}_i)^2$$
- Step 4: Compute the F-Statistic
 - If **F is significant**, at least one group mean is different.
 - **Step 5: Compare with Critical Value**
 - If **F > Critical Value**, we reject H0 and conclude that at least one teaching method is different.

Imp Note

- ANOVA and t-tests check if multiple groups are different, so they assume equality first.
 - Z-tests and Chi-square tests usually test a specific claim or association, so we set H_0 as the claim itself.
-
- ANOVA: "Let's assume no difference and see if data forces us to reject this idea."
 - Z-test/Chi-square: "Let's assume the claim is correct and check if data supports it."

Imp Note

- The decision to use a **t-test** or a **z-test** depends on two key factors:

1. Sample Size (n)

1. If $n \geq 30$, use the **z-test** (Central Limit Theorem applies).
2. If $n < 30$, use the **t-test** (small sample size).

2. Population Standard Deviation (σ) Known?

1. If the **population standard deviation (σ) is known**, use the **z-test**.
2. If **σ is unknown** and we only have the **sample standard deviation (s)**, use the **t-test**.

Imp Note Chi Square test

- **1. Chi-Square Test (for Categorical Data)**
- The **Chi-Square test** is used when we are dealing with **categorical (nominal) variables** and checking for associations or distributions.
- **Chi-Square Goodness-of-Fit Test** → Checks if a sample follows a given distribution.
- **Chi-Square Test of Independence** → Checks if two categorical variables are related.
- **Use Chi-Square when:**
 - ✓ Data is **categorical** (e.g., gender, payment method, customer satisfaction levels).
 - ✓ You want to test if categories are **independent** (e.g., Are gender and product preference related?).
- **Example for Chi-Square Test of Independence:**
- Suppose we have survey data on **payment method (Credit Card, UPI, Cash)** and **Gender (Male, Female)**.
- We want to check if payment method preference **depends on gender**.
- Since both variables are categorical, we use the **Chi-Square Test of Independence**.

Imp Note ANOVA

- **ANOVA (Analysis of Variance) (for Comparing Multiple Means)**
- **ANOVA** is used when we are comparing the **means of three or more groups** to see if at least one is significantly different.

Use ANOVA when:

- ✓ The dependent variable is **continuous (numerical)**.
- ✓ The independent variable is **categorical with 3 or more groups**.
- ✓ You want to check if there is a **significant difference among group means**.
- **Example for One-Way ANOVA:**
- Suppose we have **three different teaching methods (A, B, C)** and their **average test scores**.
- We want to test if the **mean test scores are the same across all methods**.
- Since we are comparing **more than two means**, we use **One-Way ANOVA**.