Business Statistics

Dr. Arjun Anil Kumar

National Institute of Technology Calicut
arjunanilk@nitc.ac.in

November 9, 2025

Presentation Overview

1. Introduction
   - Variables
   - Sampling
     - Probabilistic Sampling
     - Non-Probabilistic Sampling

2. Parameter Estimation
   - Point Estimation
   - Interval Estimation
   - Hypothesis Testing

Presentation Overview

3. Statistical Tests
   - Test for Means
     - One Group Tests
     - Two Group Tests
   - Test for Variance
     - One Group Test
     - Two Group Tests
   - Test for Proportion
     - One Group Test
     - Two Group Test
     - Multiple Group Test
4. Linear Regression
5. ANOVA - Analysis of Variance

Introduction

Variables & Sampling Techniques

Variable

Variable

A variable is a measurable characteristic that can assume different values among subjects or instances (e.g., age, height, income, gender).

Variables are of different types -

- Nominal
- Ordinal
- Ratio
- Interval

- **Nominal** variables represent categories with no natural order.
  For eg : Gender (Male/Female/Other), Blood Group (A, B, AB, O).
- **Ordinal** variables represent categories with a meaningful order, but no fixed interval between them.
  For eg : Education level (Primary, Secondary, Graduate), Satisfaction scale (Low, Medium, High).
- **Interval** variables represent quantitative measurements, where differences are meaningful, but absolute zero is not defined for the variable.
  For eg : Temperature in Celsius or Fahrenheit, SAT scores 400-1600
- **Ratio** variables represent quantitative measurements, where differences are meaningful and absolute zero is defined for the variable.
  For eg : Height, Weight, Income, Age and Distance.

Sampling

Sampling

The process of selecting a subset (sample) from a population to draw conclusions about the entire population. The goal of sampling is to obtain representative data in a cost-effective and time-saving way.

Sampling is of different types

- **Probabilistic or Random Sampling**
- **Non-probabilistic Sampling**

Probabilistic Sampling

Every element of the population has a known, non-zero probability of being included in the sample.

- **Simple Random Sampling** : Every unit has an equal chance.
  For eg : Lottery draw, random number generator.
- **Systematic Sampling** : Select every k-th element after a random start.
  For eg : Survey every 10th customer entering a mall.
- **Stratified Sampling** : Divide population into strata (groups), then sample randomly within each.
  For eg : Divide students by year (UG, PG, PhD) and sample proportionally.
- **Cluster Sampling** : Divide population into clusters (often geographically), then randomly choose some clusters and study all units inside them.
  For eg : Select 5 villages randomly and survey all households within them.

Random Sampling with and without Replacement

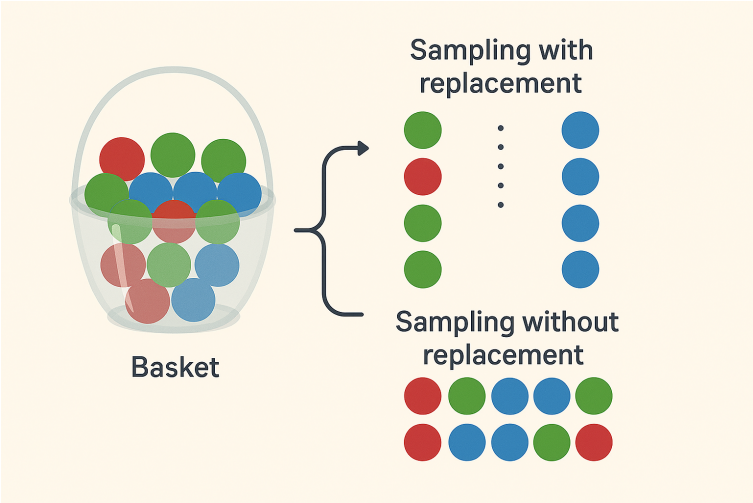


Figure: Random Sampling with and without Replacement

Stratified Sampling

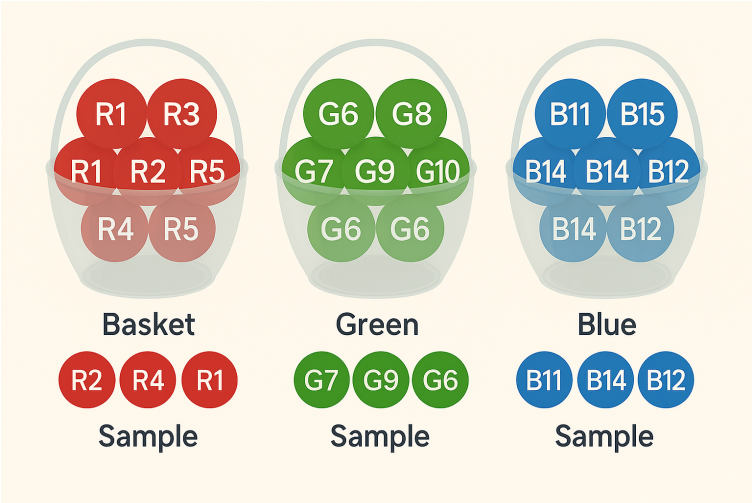


Figure: Stratified Sampling with Strata's of Similar Size

Non-Probabilistic Sampling

Not all elements have a chance of being selected; depends on researcher's choice/judgment.

- **Convenience Sampling** : Choose whoever is easiest to reach.
  For eg : Interviewing people outside your office.
- **Purposive (Judgmental) Sampling** : Researcher selects units based on specific purpose/criteria.
  For eg : Selecting only expert farmers for a study.
- **Quota Sampling** : Similar to stratified, but selection within strata is non-random.
  For eg : Interview 60 UG and 40 PG shoppers, chosen conveniently
- **Snowball Sampling** : Existing participants recruit future participants.
  For eg : Studying rare diseases or hidden communities.

When to Use Each Sampling Technique

Sampling	When to Use
Simple Random	For small and homogeneous population
Systematic	When a population list is available and ordered
Stratified	When population has clear subgroups (strata) and you want proportional representation from each subgroup.
Cluster	Large, geographically spread population where direct sampling of individuals is costly or difficult
Convenience	For quick, easy, and low-cost data collection; suitable for exploratory or pilot studies (not representative).
Purposive	When studying experts, special cases, or specific groups of interest (common in qualitative research).
Quota	When subgroup proportions are needed but random sampling isn't feasible; often used in market research.
Snowball	When the population is hidden, rare, or hard-to-reach, and existing participants can help recruit others.

Parameter Estimation

Point Estimation, Interval Estimation & Hypothesis Testing

Parameter Estimation

- Parameter Estimation means estimating the parameters of the population distribution.
- Parameter Estimation has two methods
  - Point Estimation - Estimates the parameter as a single point or value. Examples are
    Single point estimate for population mean is sample mean
    Single point estimate for population variance is sample variance
    Single point estimate for population proportion is sample proportion
  - Interval Estimation - Estimates the interval with lower and upper limits for a parameter. Examples are
    Interval Estimates for population mean with a upper and lower limit
    Interval Estimates for population variance with a upper and lower limit
    Interval Estimates for population proportion with a upper and lower limit

Point Estimation

Unbiased, Consistent and Efficient Estimators

Definition

A **statistic** is a numerical value calculated from sample data that is used to describe or summarize some characteristic of the sample.

Formally, if a random sample is $X_1, X_2, \ldots, X_n$, then a statistic is any function of the sample observations:

$$T = f(X_1, X_2, \ldots, X_n)$$

- Sample Mean: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- Sample Variance: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$
- Sample Proportion: $\hat{p} = \frac{k}{n}$, where k is the number of successes, n is the sample size.

A **Parameter** is a numerical characteristic of the population (e.g., μ, σ^2, p) and a **Statistic** is a numerical characteristic of the sample (e.g., $\bar{x}, s^2, \hat{p}$).

- Sample Difference of Means:

$$\bar{X}_1 - \bar{X}_2 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_{1i} - \frac{1}{n_2}\sum_{i=1}^{n_2} X_{2i}$$

- Sample Difference of Proportions:

$$\hat{p}_1 - \hat{p}_2 = \frac{k_1}{n_1} - \frac{k_2}{n_2},$$

where k_1 and k_2 are the number of successes in samples of size n_1 and n_2, respectively.

Definition

A **point estimator** is a statistic used to provide a **single best estimate** of an unknown population parameter. The value obtained from the sample is called a **point estimate**.

Population Parameter	Point Estimator (Sample Statistic)
Mean (μ)	Sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
Variance (σ^2)	Sample variance $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$
Proportion (p)	Sample proportion $\hat{p} = \frac{k}{n}$

Point Estimation

- A point estimator is a **random variable** because it depends on the sample.
- Desirable properties:
  - **Unbiasedness:** Expected value equals the population parameter
  - **Consistency:** Estimate approaches true parameter as sample size increases
  - **Efficiency:** Smaller variance among unbiased estimators

Unbiased Estimator

Definition

An estimator $\hat{\theta}$ of a population parameter θ is said to be **unbiased** if its expected value equals the true parameter:

$$E[\hat{\theta}] = \theta$$

If $E[\hat{\theta}] \neq \theta$, the estimator is **biased**, and the difference is called **bias**:

$$\text{Bias}(\hat{\theta}) = E[\hat{\theta}] - \theta$$

Parameter	**Estimator**	**Unbiased?**
Mean (μ)	Sample mean $\bar{X}$	Yes, $E[\bar{X}] = \mu$
Variance (σ^2)	Sample variance s^2	Yes, $E[s^2] = \sigma^2$
Proportion (p)	Sample proportion $\hat{p}$	Yes, $E[\hat{p}] = p$

Consistent Estimator

Definition

An estimator $\hat{\theta}_n$ of a population parameter θ is said to be **consistent** if it converges in probability to the true parameter as the sample size $n \to \infty$:

$$\hat{\theta}_n \xrightarrow{P} \theta \quad \text{as } n \to \infty$$

Intuitively: as the sample size increases, the estimator gets **closer and closer** to the true parameter.

Note : There are biased and unbiased estimators that are consistent.

Efficient Estimator

Definition

An estimator $\hat{\theta}$ of a population parameter θ is said to be **efficient** if it is **unbiased** and has the **smallest variance** among all unbiased estimators of θ:

$$\text{Var}(\hat{\theta}) \leq \text{Var}(\tilde{\theta}) \quad \text{for any other unbiased estimator } \tilde{\theta}.$$

Efficiency measures how **precisely** an estimator estimates the parameter.

- For a normal population $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, the **sample mean** $\bar{X}$ is an efficient estimator of μ.
- Among all unbiased estimators of μ, $\bar{X}$ has the minimum variance σ^2/n.

$$\text{Relative Efficiency of } \hat{\theta}_1 \text{ to } \hat{\theta}_2 = \frac{\text{Var}(\hat{\theta}_2)}{\text{Var}(\hat{\theta}_1)}$$

Step 1: Identify estimators

$$\hat{\theta}_1 = \text{sample mean } \bar{X}, \quad \hat{\theta}_2 = \text{sample median}$$

Step 2: Variances for normal distribution

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n}, \quad \text{Var(median)} \approx \frac{\pi\sigma^2}{2n} \quad \text{(for large } n\text{)}$$

Step 3: Compute relative efficiency

$$\text{RE(mean, median)} = \frac{\text{Var(median)}}{\text{Var}(\bar{X})} = \frac{\frac{\pi\sigma^2}{2n}}{\frac{\sigma^2}{n}} = \frac{\pi}{2} \approx 1.57$$

Relative efficiency of 1.57 means the sample mean achieves the same precision with fewer observations than the sample median.

Interval Estimation

Interval Estimation

Definition (Interval Estimation)

Interval estimation refers to estimating a population parameter (like a mean or proportion or variance) using a range of plausible values instead of a single point estimate.

$$\text{Confidence Interval (CI)} = \text{Point Estimate (PE)} \pm \text{Margin of Error (ME)}$$

The confidence interval of a population parameter is an interval between PE-ME to PE+ME.

- Assume the sample mean is $\hat{\mu}$ = 1000
- The sample mean does not say anything about how close the sample mean $\hat{\mu}$ is to the population mean μ
- Is the population mean likely to be in the range 900 and 1000 or $900 \leq \mu \leq 1100$
- Is the population mean likely to be in the range 990 to 1010 or $990 \leq \mu \leq 1010$?
- An interval estimate for a population parameter is called a **confidence interval**.
- Shorter the length of the interval, more precise is the estimation.

- Suppose that $X_1, X_2, .., X_n$ is a random sample from a normal distribution with unknown mean μ and known variance σ^2.
- CLT says

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \tag{1}$$

  As $n- > \infty$, is the standard normal distribution.
- A confidence interval estimate for μ is an interval of the form $l \leq \mu \leq u$, where the end points l and u are computed from the sample data.
- Because different samples will produce different values of l and u, these end-points are values of random variables L and U, respectively.

- Is it possible to determine the values of L and U such that the following probability condition is satisfied. $P\left[L \leq \mu \leq U\right] = 1 - \alpha$, where 1 - α is called the confidence.
- The answer is yes! $P\left[-z_{\frac{\alpha}{2}} \leq Z \leq z_{\frac{\alpha}{2}}\right] = 1 - \alpha$
  where $Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$
- For a sample of size n, the confidence interval for population mean μ and population standard deviation σ with a confidence of 1 - α, where α is the level of significance, is given by

$$P\left[\bar{X} - ME \leq \mu \leq \bar{X} + ME\right] = 1 - \alpha \tag{2}$$

$$ME = z_{\frac{\alpha}{2}} * SE \tag{3}$$

$$SE = \frac{\sigma}{\sqrt{n}} \tag{4}$$

$$P\left[\bar{X} - z_{\frac{\alpha}{2}}.\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\frac{\alpha}{2}}.\frac{\sigma}{\sqrt{n}}\right] = 1 - \alpha$$

The confidence interval for population mean μ is defined as $\left[\bar{X} \pm z_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right]$.

The width of the confidence interval represents the precision of the estimate.

- The width of the confidence interval (Upper Limit - Lower Limit) is given by 2ME or $2.z_{\frac{\alpha}{2}} * SE$ or $2.z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}$
- When confidence (1-α) is increased, $Z_{\frac{\alpha}{2}}$ is increased. For example : For 95% confidence or α =.05, we have $Z_{.025}$ = 1.96 and 99% confidence or α = .1 or $Z_{.005}$ = 2.58.
- Confidence level ↑ Confidence Interval ↑ Precision ↓

For population mean $\mu = 50$, $\sigma = 10$ and n = 30, $z_{.025} = 1.96$, the 95% confidence intervals are shown below.

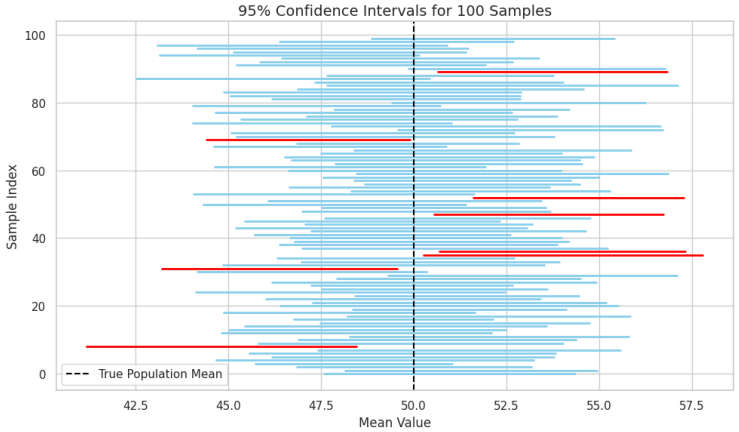


Figure: Confidence Interval for Population Mean

- Ideally you would want a high precision and high confidence interval.
- But when we increase the confidence level or the confidence interval, the precision is actually decreasing.
- **But for a given confidence interval, we can improve the precision by increasing the sample size n**
- The sampling error is the difference between a sample statistic and the true population parameter it estimates. Example $\bar{X} - \mu$ or $\hat{p} - p$

$$P\left[\bar{X} - z_{\frac{\alpha}{2}}.\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\frac{\alpha}{2}}.\frac{\sigma}{\sqrt{n}}\right] = 1 - \alpha \tag{5}$$

$$P\left[-z_{\frac{\alpha}{2}} \leq \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{\frac{\alpha}{2}}\right] = 1 - \alpha \tag{6}$$

$$P\left[|E| < z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right] = 1 - \alpha \tag{7}$$

$$P\left[|E| < E_{max}\right] = 1 - \alpha \tag{8}$$

If $\bar{X}$ is the sample mean and μ is the population mean, we can be $100(1-\alpha)$ sure that the sampling error $\bar{X} - \mu$ will not exceed an error E_{max} if the sample size n = $\left[\frac{z_{\frac{\alpha}{2}} * \sigma}{E_{max}}\right]^2$

- Error $E_{max} \downarrow$ implies increase in the sample size n if the confidence level $(1-\alpha)$ and population standard deviation σ is fixed.
- As population standard deviation $\sigma \downarrow$ the sample size n also reduces when the error E and the confidence level $(1-\alpha)$ is fixed.
- As the level of confidence increases or $z_{\frac{\alpha}{2}}$ increases, the sample size n also increases when the error E and population standard deviation σ is fixed.

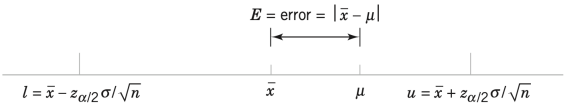


Figure: Error in estimating μ with $\bar{X}$

- Suppose that $X_1, X_2, .., X_n$ is a random sample from a normal distribution with unknown mean μ and unknown variance σ^2.
- CLT says

$$Z = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \tag{9}$$

As $n- > \infty$, is the standard normal distribution.

$$P\left[\hat{\mu} - z_{\frac{\alpha}{2}}.\frac{s}{\sqrt{n}} \leq \mu \leq \hat{\mu} + z_{\frac{\alpha}{2}}.\frac{s}{\sqrt{n}}\right] = 1 - \alpha \tag{10}$$

- Suppose that $X_1, X_2, .., X_n$ is a random sample from a normal distribution with unknown mean μ and unknown variance σ^2.
- CLT says

$$T = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \tag{11}$$

As $n- > \infty$, is the T distribution with n-1 degrees of freedom

$$P\left[\hat{\mu} - t_{\frac{\alpha}{2},n-1}.\frac{s}{\sqrt{n}} \leq \mu \leq \hat{\mu} + t_{\frac{\alpha}{2},n-1}.\frac{s}{\sqrt{n}}\right] = 1 - \alpha \tag{12}$$

- We need to construct confidence interval for population proportion.
- Assume we randomly sample x_1,x_2,x_3,..x_n
- Note that the samples x_1,x_2,x_3,..x_n are drawn from a bernouli distribution of parameter p.
- The sample proportion is the point estimator of unknown population population proportion p.

$$\hat{p} = \frac{\sum_{i=1}^{n}[x_i]}{n} \tag{13}$$

- $\hat{p}$ is a random variable with mean $\frac{1}{n} * n * p = p$ and variance $\frac{1}{n^2}.n.p.(1-p) = \frac{p(1-p)}{n}$.
- As per CLT, if n is large, the distribution of Z approaches $\mathcal{N}(0,1)$

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p.(1-p)}{n}}} \tag{14}$$

- $$P\big[-z_{\frac{\alpha}{2}} < Z < z_{\frac{\alpha}{2}}\big] = 1-\alpha \tag{15}$$
- $$P\big[\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{p.(1-p)}{n}} \leq p < \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{p.(1-p)}{n}}\big] = 1-\alpha \tag{16}$$
- Confidence Interval for Proportion

$$P\big[\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}.(1-\hat{p})}{n}} \leq p < \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}.(1-\hat{p})}{n}}\big] = 1-\alpha \tag{17}$$

- Choice of sample size n based on max error E_{max} or $max|(\hat{p} - p)|$

$$P\left[|E| < E_{max}\right] = 1 - \alpha \tag{18}$$

If $\hat{p}$ is the sample proportion and p is the population proportion, we can be $100(1-\alpha)$, that the error $|\hat{p} - p|$ will not exceed an error E_{max} if the sample size n = $\left[\frac{z_{\frac{\alpha}{2}} * \sigma}{E_{max}}\right]^2$ or $\left[\frac{z_{\frac{\alpha}{2}} * \sqrt{p.(1-p)}}{E_{max}}\right]^2$

Out of random sample of 85 students, 10 students voted for party A and 75 voted for the rest. Compute the confidence interval for the population proportion voting for party A
Sample proportion $\hat{p} = \frac{10}{85}$ For a two sided 95% interval, $\alpha = .05$, $z_{.025} = 1.96$
The confidence interval of population proportion is given by
$[\hat{p} - z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}.(1-\hat{p})}{n}} \leq p < \hat{p} + z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}.(1-\hat{p})}{n}}]$
.05 <= p <= .19
How many samples do we pick for taking a decision assuming the max error $max|(\hat{p} - p)|$ is .05?
$n = \left[\frac{z_{\frac{\alpha}{2}} * \sqrt{p.(1-p)}}{E_{max}}\right]^2$

For observed sample, sample size n = $\left[\frac{z_{\frac{\alpha}{2}} * \sqrt{\hat{p}.(1-\hat{p})}}{E_{max}}\right]^2$ = 163

If no observed sample is there, maximum sample size n = $\left[\frac{z_{\frac{\alpha}{2}} * \sqrt{.5*(1-.5)}}{E_{max}}\right]^2$ = 385

In the Framingham Offspring study 1,219 subjects were on anti-hypertensive medication out of 3,532 total subjects. Construct the confidence interval on the population proportion with 95% confidence?

- $\hat{p} = \frac{1219}{3532}$
- Confidence interval equals to [$\hat{p}$ - E,$\hat{p}$+E]
- Error E is $z_{\frac{\alpha}{2}}$*SE($\hat{p}$)
- SE($\hat{p}$) = $\sqrt{\frac{p.(1-p)}{n}}$
- Since p is unknown, SE($\hat{p}$) = $\sqrt{\frac{\hat{p}.(1-\hat{p})}{n}}$
- n is sample size = 3532
- Confidence Interval is [.329,.361]

- If $X_1, X_2, X_3 \ldots X_n$ are standard normal variables with mean 0 and variance 1, the variable χ^2 expressed as the sum of squares of standard normal variable would have a χ^2 distribution.

$$\chi^2 = X_1^2 + X_2^2 + X_3^2 + X_n^2 \tag{19}$$

- $\chi^2_{n-1} = \frac{n-1.s^2}{\sigma^2}$ is a chi-square distribution with n-1 degrees of freedom.
- $P\left[\chi^2_{1-\frac{\alpha}{2},n-1} \leq \chi^2 \leq \chi^2_{\frac{\alpha}{2},n-1}\right] = 1-\alpha$
- $\chi^2_{1-\frac{\alpha}{2},n-1} \leq \frac{n-1.s^2}{\sigma^2} \leq \chi^2_{\frac{\alpha}{2},n-1}$
- The confidence interval of σ^2 is $\left[\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2},n-1}}, \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}\right]$

Parameter	Interval Type	Distribution Used
Mean (μ)	$\bar{X} \pm z_{\alpha} \frac{s}{\sqrt{n}}$	t-distribution , n>40
Mean (μ)	$\bar{X} \pm t_{\alpha/2,n-1} \frac{s}{\sqrt{n}}$	t-distribution , n<40
Proportion (p)	$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	Normal (z)
Variance (σ^2)	$\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}$	Chi-square

A nutritionist measures the daily vitamin C intake of a random sample of $n = 25$ adults. The sample mean is $\bar{x} = 100$ mg and the sample standard deviation is $s = 15$ mg.

1. Construct a 95% confidence interval for the population mean μ and decide whether there is evidence at the 5% level that the true mean differs from $\mu_0 = 95$ mg.

$$n = 25, \quad \bar{x} = 100, \quad s = 15, \quad \text{confidence level} = 95\%$$

$$df = n - 1 = 24, \quad t_{0.025,24} \approx 2.064$$

$$SE = \frac{s}{\sqrt{n}} = \frac{15}{\sqrt{25}} = 3$$

$$ME = t_{0.025,24} \cdot SE = 2.064 \times 3 \approx 6.192$$

$$\bar{x} \pm ME = 100 \pm 6.192 \Rightarrow (93.808, 106.192)$$

- A public-health survey of $n = 500$ adults found that $x = 65$ reported using Drug X during the past month.
- Construct a 95% confidence interval for the population proportion p of adults who use Drug X and check whether the null hypothesis $H_o : p = .12$ can be rejected.

$$\hat{p} = \frac{x}{n} = \frac{65}{500} = 0.13$$

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.13 \times 0.87}{500}} \approx 0.01504$$

$$ME = z_{\alpha/2} \times SE(\hat{p}) = 1.96 \times 0.01504 = 0.02948$$

$$\hat{p} \pm ME = 0.13 \pm 0.02948$$

$$\Rightarrow (0.1005,\ 0.1595)$$

Hypothesis Testing

Test for Mean, Proportion and Variance
Single Group | Two Group | Multiple Groups

Hypothesis Testing

A hypothesis is an assumption or statement about a population parameter (such as mean, variance, or proportion) that we test using sample data.

- Groups - Single Group | Two Group | Multiple Groups
- Samples - Dependent (Paired) Samples | Independent Samples
- Assumptions - Parameteric | Non-parametric Tests

Hypothesis Tesing

- Hypothesis for Population Mean - One Group
- Hypothesis for Population Mean - Two Group
- Hypothesis for Population Mean - Multiple Group
- Hypothesis for Population Proportion - One Group
- Hypothesis for Population Proportion - Two Group
- Hypothesis for Population Proportion - Multiple Group
- Hypothesis for Population Variance - One Group
- Hypothesis for Population Variance - Two Group
- Hypothesis for Population Variance - Multiple Group

One-Mean Hypothesis

A hypothesis is an assumption or statement about a population mean that we test using sample data.

- **Null Hypothesis** (H_0): The default assumption, states no effect or no difference.

$$H_0 : \mu = \mu_0$$

- **Alternative Hypothesis** (H_1): Competes with H_0, states that an effect or difference exists.

$$\text{Two-tailed: } H_1 : \mu \neq \mu_0$$
$$\text{Right-tailed: } H_1 : \mu > \mu_0$$
$$\text{Left-tailed: } H_1 : \mu < \mu_0$$

One-Proportion Hypothesis

A hypothesis is an assumption or statement about a population proportion that we test using sample data.

- **Null Hypothesis** (H_0): The default assumption, states no effect or no difference.

$$H_0 : p = p_0$$

- **Alternative Hypothesis** (H_1): Competes with H_0, states that an effect or difference exists.

$$\text{Two-tailed: } H_1 : p \neq p_0$$
$$\text{Right-tailed: } H_1 : p > p_0$$
$$\text{Left-tailed: } H_1 : p < p_0$$

One-Variance Hypothesis

A hypothesis is an assumption or statement about a population variance that we test using sample data.

- **Null Hypothesis** (H_0): The default assumption, states no effect or no difference.

$$H_0 : \sigma^2 = \sigma_0^2$$

- **Alternative Hypothesis** (H_1): Competes with H_0, states that an effect or difference exists.

$$\text{Two-tailed: } H_1 : \sigma^2 \neq \sigma_0^2$$
$$\text{Right-tailed: } H_1 : \sigma^2 > \sigma_0^2$$
$$\text{Left-tailed: } H_1 : \sigma^2 < \sigma_0^2$$

Two-Mean Hypothesis

A hypothesis is an assumption or statement about the difference between two population means that we test using sample data.

- **Null Hypothesis** (H_0): No difference between the two population means.

$$H_0 : \mu_1 = \mu_2 \quad \text{or} \quad H_0 : \mu_1 - \mu_2 = 0$$

- **Alternative Hypothesis** (H_1): A difference exists between the two population means.

$$\text{Two-tailed: } H_1 : \mu_1 \neq \mu_2$$
$$\text{Right-tailed: } H_1 : \mu_1 > \mu_2$$
$$\text{Left-tailed: } H_1 : \mu_1 < \mu_2$$

Two-Proportion Hypothesis

A hypothesis is an assumption or statement about the difference between two population proportions that we test using sample data.

- **Null Hypothesis** (H_0): No difference between the two population proportions.

$$H_0 : p_1 = p_2 \quad \text{or} \quad H_0 : p_1 - p_2 = 0$$

- **Alternative Hypothesis** (H_1): A difference exists between the two population proportions.

$$\begin{aligned} \text{Two-tailed: } & H_1 : p_1 \neq p_2 \\ \text{Right-tailed: } & H_1 : p_1 > p_2 \\ \text{Left-tailed: } & H_1 : p_1 < p_2 \end{aligned}$$

Two-Variance Hypothesis

A hypothesis is an assumption or statement about the ratio of two population variances that we test using sample data.

- **Null Hypothesis** (H_0): The two variances are equal.

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad \text{or} \quad H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$$

- **Alternative Hypothesis** (H_1): The variances differ.

$$\text{Two-tailed: } H_1 : \sigma_1^2 \neq \sigma_2^2$$
$$\text{Right-tailed: } H_1 : \sigma_1^2 > \sigma_2^2$$
$$\text{Left-tailed: } H_1 : \sigma_1^2 < \sigma_2^2$$

Multiple-Mean Hypothesis

A hypothesis is an assumption or statement about the means of multiple populations that we test using sample data.

- **Null Hypothesis** (H_0): All population means are equal.

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

- **Alternative Hypothesis** (H_1): At least one population mean is different.

$$H_1 : \text{Not all } \mu_i \text{ are equal}, \quad i = 1, \ldots, k$$

Multiple-Proportion Hypothesis

A hypothesis is an assumption or statement about the proportions of multiple populations that we test using sample data.

- **Null Hypothesis** (H_0): All population proportions are equal.

$$H_0 : p_1 = p_2 = \cdots = p_k$$

- **Alternative Hypothesis** (H_1): At least one population proportion is different.

$$H_1 : \text{Not all } p_i \text{ are equal}, \quad i = 1, \ldots, k$$

Multiple-Variance Hypothesis

A hypothesis is an assumption or statement about the variances of multiple populations that we test using sample data.

- **Null Hypothesis** (H_0): All population variances are equal.

$$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$$

- **Alternative Hypothesis** (H_1): At least one population variance is different.

$$H_1 : \text{Not all } \sigma_i^2 \text{ are equal}, \quad i = 1, \ldots, k$$

Hypothesis Testing Framework

Step 1: State the Hypotheses

- Two-tailed test:
$$H_0 : \theta = \theta_0, \quad H_1 : \theta \neq \theta_0$$
- Right-tailed test:
$$H_0 : \theta = \theta_0, \quad H_1 : \theta > \theta_0$$
- Left-tailed test:
$$H_0 : \theta = \theta_0, \quad H_1 : \theta < \theta_0$$

Step 2: Choose Significance Level (α)

- Commonly $\alpha = 0.05$ or 0.01
- α is the probability of Type I error (rejecting H_0 when true)

Hypothesis Testing Framework

Step 3: Select Test Statistic

Situation	Test Statistic
Population mean, known σ, large samples	$Z = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$
Population mean, unknown σ, small samples	$t = \frac{\bar{X}-\mu_0}{s/\sqrt{n}}$
Population proportion	$Z = \frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}}$
Population variance	$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$
Comparing two variances	$F = \frac{s_1^2}{s_2^2}$

Hypothesis Testing Framework

Step 4: Determine Rejection Region

- Two-tailed:

$$|Z| > Z_{\alpha/2}, \quad t > t_{\alpha/2} \text{ or } t < -t_{\alpha/2}, \quad \chi^2 < \chi^2_{\alpha/2,\, df} \text{ or } \chi^2 > \chi^2_{1-\alpha/2,\, df}$$

- Right-tailed:

$$Z > Z_{\alpha}, \quad t > t_{\alpha}, \quad \chi^2 > \chi^2_{\alpha,\, df}$$

- Left-tailed:

$$Z < -Z_{\alpha}, \quad t < -t_{\alpha}, \quad \chi^2 < \chi^2_{1-\alpha,\, df}$$

Step 5: Make Decision

- Reject H_0 if test statistic falls in rejection region
- Otherwise, fail to reject H_0

Statistical Tests

One Group, Two Group and Multi Groups

Test for Means

One Group & Two Group & Multi-Group Tests

One Group Tests

Z & t Tests

Test for Means

Abraham De Moivre
Z-Distribution

Willain Sealy Gosset
t-Distribution

z-Distribution

- The **z-distribution** is the **standard normal distribution**, with mean 0 and standard deviation 1:

$$Z \sim N(0, 1)$$

- Any normal variable $X \sim N(\mu, \sigma^2)$ can be **standardized** to a z-score using:

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

- The z-distribution is symmetric about 0, and the total area under the curve is 1.
- Abraham de Moivre first introduced the normal distribution while studying the binomial distribution.
- When the population variance σ^2 is known, the z-test statistic for a sample mean $\bar{X}$ is:

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

t-Distribution

- William Gosset (Student) showed that when the sample size is small, the t-statistic for a sample mean $\bar{X}$ is

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

  follows a **t-distribution** rather than a normal distribution.
- The **t-distribution** is wider and has heavier tails than the normal distribution, reflecting the increased uncertainty in $\bar{X}$ due to small n.
- The shape of the t-distribution depends on the degrees of freedom:

$$df = n - 1$$

  As n increases, the t-distribution approaches the normal distribution. Therefore, for large samples ($n \geq 30$), a z-test can be safely used.

Z-distribution and t-distribution

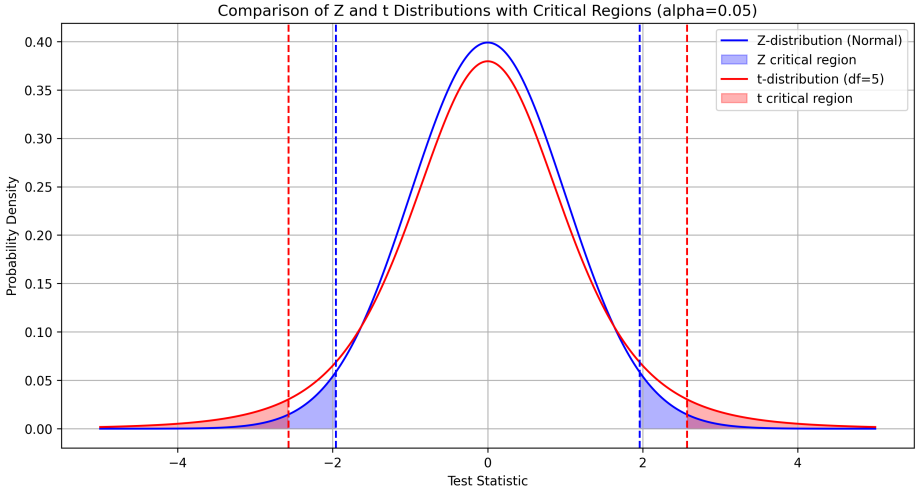


Figure: Z-distribution and t-distribution

Test for Means - One Group

- Test for Means - One Group is a parametric test as the following assumptions are made.
  - Normality : The population has a normal distribution.
  - Scale : Data should be measured at interval or ratio scale.
  - Random Sampling : Samples are randomly drawn.
  - Independent Observations: Samples are independent of each other.
- If sample size $n < 30$, use a t-test if "$X_1, X_2, \ldots, X_n \sim$ i.i.d. with mean μ, unknown σ, and approximately normal population".
- If sample size $n > 30$, use a Z-test if "$X_1, X_2, \ldots, X_n \sim$ i.i.d. with mean μ, known/unknown σ."

Hypotheses:

- **Two-tailed:**

$$H_0 : \mu = \mu_0, \quad H_1 : \mu \neq \mu_0$$

- **Right-tailed:**

$$H_0 : \mu = \mu_0, \quad H_1 : \mu > \mu_0$$

- **Left-tailed:**

$$H_0 : \mu = \mu_0, \quad H_1 : \mu < \mu_0$$

Test Statistic:

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

Decision Rule:

- Two-tailed: Reject H_0 if $|Z| > Z_{\alpha/2}$
- Right-tailed: Reject H_0 if $Z > Z_{\alpha}$
- Left-tailed: Reject H_0 if $Z < -Z_{\alpha}$

Test for Means - One Group

Small Sample ($n < 30$) – t-test

Hypotheses:

- **Two-tailed:**

$$H_0 : \mu = \mu_0, \quad H_1 : \mu \neq \mu_0$$

- **Right-tailed:**

$$H_0 : \mu = \mu_0, \quad H_1 : \mu > \mu_0$$

- **Left-tailed:**

$$H_0 : \mu = \mu_0, \quad H_1 : \mu < \mu_0$$

Test Statistic:

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}, \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2, \quad df = n - 1$$

Decision Rule:

- Two-tailed: Reject H_0 if $|t| > t_{\alpha/2,\, n-1}$
- Right-tailed: Reject H_0 if $t > t_{\alpha,\, n-1}$
- Left-tailed: Reject H_0 if $t < -t_{\alpha,\, n-1}$

t-Statistic, Z-Statistic & χ^2 Statistic

Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. Define

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Standard normal or Z-Statistic

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

Chi-square χ^2 Statistic U

$$U = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

t-Statistic, Z-Statistic & χ^2 Statistic

t-statistic

$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{(\bar{X} - \mu)/(\sigma/\sqrt{n})}{S/\sigma} = \frac{Z}{S/\sigma}.$$

Since

$$\frac{S}{\sigma} = \sqrt{\frac{(n-1)S^2}{\sigma^2} \cdot \frac{1}{n-1}} = \sqrt{\frac{U}{n-1}},$$

we obtain

$$\boxed{t = \frac{Z}{\sqrt{U/(n-1)}}}$$

where $Z \sim N(0,1)$ and $U \sim \chi^2_{n-1}$.

- By **Cochran's theorem**:

$$\underbrace{\sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2}}_{\chi^2_n} = \underbrace{\frac{n(\bar{X} - \mu)^2}{\sigma^2}}_{\chi^2_1} + \underbrace{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2}}_{\chi^2_{n-1}},$$

  and these two components are independent.
- The first term corresponds to Z^2, the second to U.
- Hence Z and U are independent.

t-Statistic, Z-Statistic & χ^2 Statistic

Because $Z \sim N(0,1)$ and $U \sim \chi^2_{n-1}$ are independent,

$$T = \frac{Z}{\sqrt{U/(n-1)}} \sim t_{n-1}.$$

A Student-t distributed variable t with $\nu = n-1$ degrees of freedom has a probability density function $f_T(t)$

$$f_T(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \qquad \nu = n-1.$$

For any complex number z with $\Re(z) > 0$,

$$\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}\,dx$$

$$\Gamma(z+1) = z\,\Gamma(z)$$

Mean and Variance of Students t Distribution

For a t-distribution with ν degrees of freedom,

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \quad -\infty < t < \infty$$

Mean

$$E[t] = \begin{cases} 0, & \nu > 1, \\ \text{undefined}, & \nu \leq 1. \end{cases}$$

Variance

$$\mathrm{Var}(t) = \begin{cases} \dfrac{\nu}{\nu-2}, & \nu > 2, \\ \infty, & 1 < \nu \leq 2, \\ \text{undefined}, & \nu \leq 1. \end{cases}$$

Two Group Tests

Independent Sample Tests & Dependent Sample Tests

Applicable to:

- Independent two-sample tests (Z-test or t-test)
- Dependent (paired) sample t-test

General Assumptions:

- **Normality:** Each population from which samples are drawn is normally distributed.
- **Scale:** Data are measured on an interval or ratio scale.
- **Random Sampling:** Samples are randomly and independently drawn from the populations.

For Independent Two-Sample Tests:

- The two samples are independent of each other.
- Population variances are:
  - **Equal** for **pooled t-test** (assume homogeneity of variances).
  - **Unequal** for **unpooled (Welch's) t-test**.
- Sample sizes can be equal or unequal.

For Paired Sample Test:

- Observations are paired (before–after, matched pairs, or repeated measures).
- Differences between pairs are normally distributed.

Independent Samples:

- Two groups are selected independently; observations in one group are not related to those in the other.
- Examples:
  - Comparing average test scores of students from two different schools.
  - Comparing mean income levels of urban and rural households.
  - Measuring the effectiveness of two different fertilizers on separate plots of land.
  - Comparing mean weights of two distinct animal species.

Dependent (Paired) Samples:

- Observations are paired — the same or matched subjects are measured twice or under two conditions.
- Examples:
  - Comparing blood pressure of patients *before* and *after* taking medication - Before/After pairs
  - Measuring students' performance *before* and *after* a training program - Before/After pairs
  - Comparing yields of the same crop under two different irrigation methods on the same plots - Matched Pairs
  - Comparing accuracy of two teaching methods on the same group of students - Matched Pairs

Hypotheses:

- **Two-tailed:**
$$H_0 : \mu_1 = \mu_2, \quad H_1 : \mu_1 \neq \mu_2$$
- **Right-tailed:**
$$H_0 : \mu_1 = \mu_2, \quad H_1 : \mu_1 > \mu_2$$
- **Left-tailed:**
$$H_0 : \mu_1 = \mu_2, \quad H_1 : \mu_1 < \mu_2$$

The difference in sample means, $\bar{X}_1 - \bar{X}_2$, is a random variable with mean $E[\bar{X}_1 - \bar{X}_2] = \mu_1 - \mu_2$ and variance $\text{Var}(\bar{X}_1 - \bar{X}_2) = \dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}$.

Test Statistic:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where

$$S_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}, \quad df = n_1 + n_2 - 2$$

Decision Rule:

- Two-tailed: Reject H_0 if $|t| > t_{\alpha/2,\, n_1+n_2-2}$
- Right-tailed: Reject H_0 if $t > t_{\alpha,\, n_1+n_2-2}$
- Left-tailed: Reject H_0 if $t < -t_{\alpha,\, n_1+n_2-2}$

Measuring the Effect Size with Cohen's d

Cohen's d: Effect Size Cohen's d measures the **standardized difference** between two means.

$$d = \frac{\bar{X}_1 - \bar{X}_2}{s_p}$$

where the pooled standard deviation is

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Interpretation:

- $d = 0.2$: Small effect — difference is minor
- $d = 0.5$: Medium effect — noticeable difference
- $d = 0.8$: Large effect — substantial difference
- $d > 1.2$: Very large effect — very substantial difference

Pooled t-test for Independent Samples (Equal Variances Assumed)

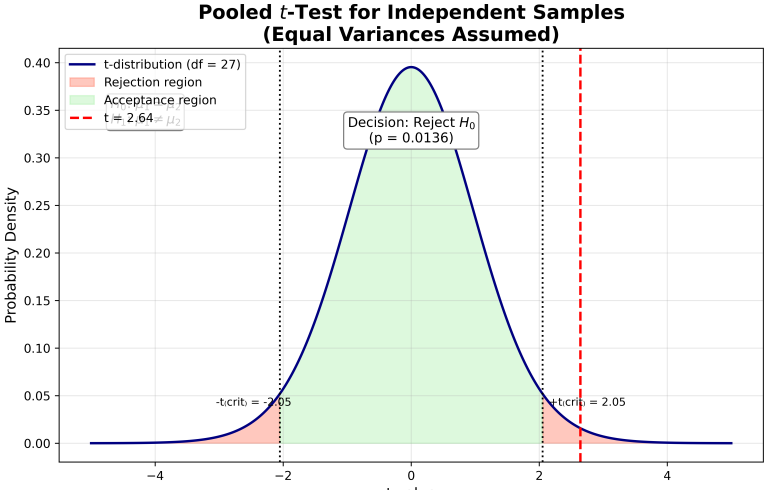
Problem: An education researcher wants to determine whether a new teaching method improves student performance compared to the traditional method. Two independent groups of students were randomly assigned to each teaching method, and their final exam scores were recorded.

$$\text{New Method:} \quad n_1 = 15,\ \bar{x}_1 = 78.4,\ s_1 = 6.2$$
$$\text{Traditional Method:} \quad n_2 = 14,\ \bar{x}_2 = 72.5,\ s_2 = 5.8$$

Assume that population variances are equal. Test at $\alpha = 0.05$ whether the new teaching method leads to a significantly different mean score.

Pooled t-test for Independent Samples (Equal Variances Assumed)

Solution



Hypotheses:

- **Two-tailed:**
$$H_0 : \mu_1 = \mu_2, \quad H_1 : \mu_1 \neq \mu_2$$
- **Right-tailed:**
$$H_0 : \mu_1 = \mu_2, \quad H_1 : \mu_1 > \mu_2$$
- **Left-tailed:**
$$H_0 : \mu_1 = \mu_2, \quad H_1 : \mu_1 < \mu_2$$

The difference in sample means, $\bar{X}_1 - \bar{X}_2$, is a random variable with mean $E[\bar{X}_1 - \bar{X}_2] = \mu_1 - \mu_2$ and variance $\text{Var}(\bar{X}_1 - \bar{X}_2) = \dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}$.

Test Statistic:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Degrees of Freedom (Welch-Satterthwaite Approximation):

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

Decision Rule:

- Two-tailed: Reject H_0 if $|t| > t_{\alpha/2,\, df}$
- Right-tailed: Reject H_0 if $t > t_{\alpha,\, df}$
- Left-tailed: Reject H_0 if $t < -t_{\alpha,\, df}$

Unpooled *t*-test for Independent Samples (Unequal Variances)

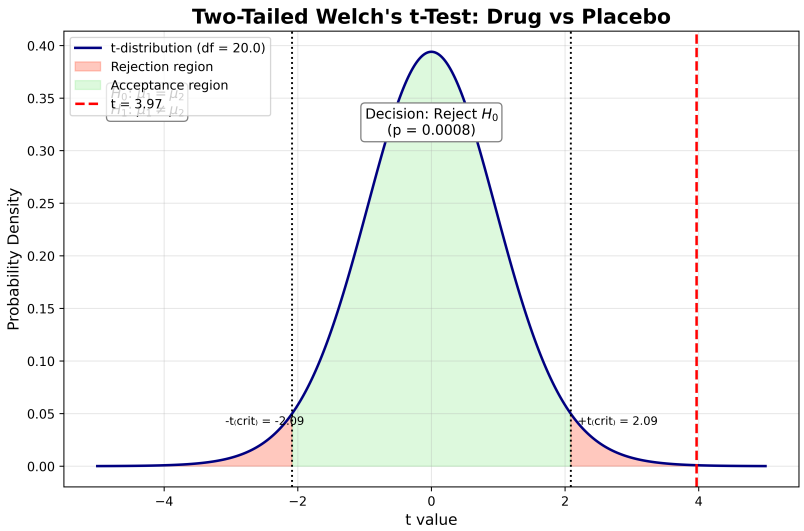
Problem: A researcher tests whether a new drug reduces symptom severity more effectively than a placebo.

$$\text{Drug group:} \quad n_1 = 12,\ \bar{x}_1 = 8.5,\ s_1 = 2.1$$
$$\text{Placebo group:} \quad n_2 = 10,\ \bar{x}_2 = 5.2,\ s_2 = 1.8$$

Test at $\alpha = 0.05$ whether the drug produces a significantly different mean reduction. Assume that the population variances are very different.

Unpooled *t*-test for Independent Samples (Unequal Variances)

Solution



- **Paired Observations:** Each observation in one group is paired with a specific observation in the other group.
- **Normality of Differences:** The differences $D_i = X_i - Y_i$ are approximately normally distributed.
- **Scale of Measurement:** Data should be interval or ratio scale.
- **Random Sampling:** Pairs are randomly selected from the population.
- **Independence of Pairs:** Each pair is independent of other pairs.

Hypotheses:

- Two-tailed: $H_0 : \mu_D = 0, \quad H_1 : \mu_D \neq 0$
- Right-tailed: $H_0 : \mu_D = 0, \quad H_1 : \mu_D > 0$
- Left-tailed: $H_0 : \mu_D = 0, \quad H_1 : \mu_D < 0$

Difference per pair: $D_i = X_i - Y_i, \quad i = 1, \ldots, n$

Sample mean and variance of differences:

$$\bar{D} = \frac{1}{n}\sum_{i=1}^{n} D_i, \quad s_D^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \bar{D})^2$$

Test Statistic:

$$t = \frac{\bar{D} - \mu_D}{s_D/\sqrt{n}}, \quad df = n - 1$$

Decision Rule:

- Two-tailed: reject H_0 if $|t| > t_{\alpha/2,n-1}$
- Right-tailed: reject H_0 if $t > t_{\alpha,n-1}$
- Left-tailed: reject H_0 if $t < -t_{\alpha,n-1}$

Paired t-test for Dependent Samples

Problem:
A researcher wants to determine whether an 8-week gym training program significantly improves participants' strength. Ten individuals are measured for their 1-repetition maximum (1RM) bench press **before** and **after** the program. If someone's 1RM on the bench press is 100 kg, it means they can lift 100 kg once

$$\begin{aligned} \text{Before training:} \quad & [100, 105, 98, 110, 102, 95, 108, 101, 99, 104] \\ \text{After training:} \quad & [105, 108, 100, 115, 106, 98, 112, 103, 101, 108] \end{aligned}$$

At a significance level of $\alpha = 0.05$, test whether the training program leads to a significant improvement in mean bench press strength.

Paired *t*-test for Dependent Samples

Solution

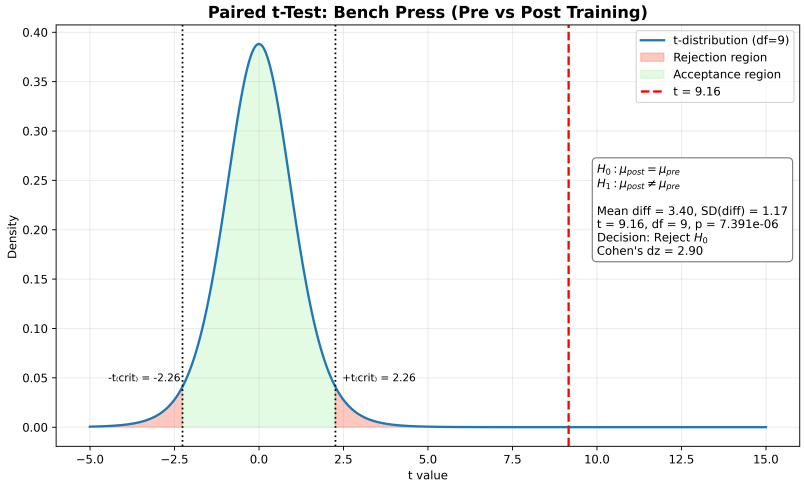


Figure: Paired t-test

Test for Variance

χ^2 & F Tests

One Group Test

χ^2 Test

Chi-square Distribution

Definition

A Chi-square random variable with n degrees of freedom is defined as

$$\chi^2_n = \sum_{i=1}^{n} Z_i^2, \quad Z_i \sim N(0,1) \text{ i.i.d.}$$

Idea of Simulation

- Generate n independent samples from $N(0,1)$.
- Square them and sum to obtain one χ^2_n sample.
- Repeat this many times to approximate the distribution.

Simulation of Chi-square Distribution

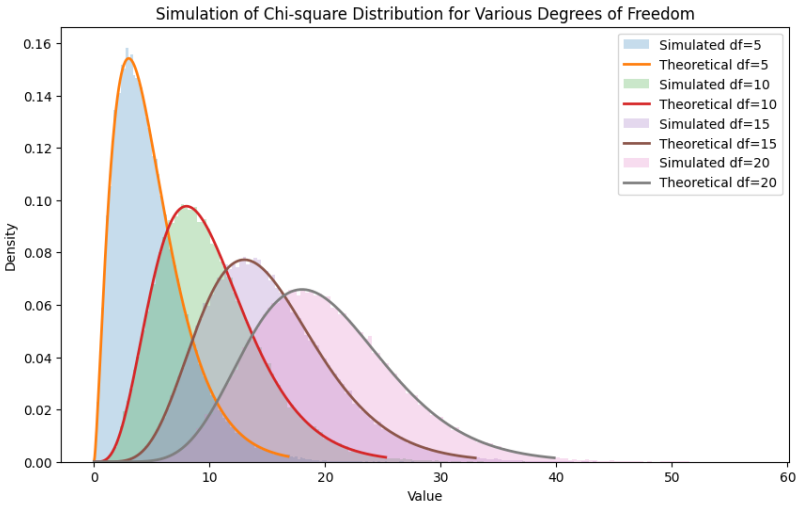


Figure: Simulation of Chi-square Distribution for varying degrees of freedom

- Test for Variance - One Group is a parametric test used to check whether the population variance equals a hypothesized value.
- **Assumptions:**
  - Normality: The population from which the sample is drawn is normally distributed.
  - Scale: Data should be measured at an interval or ratio scale.
  - Random Sampling: Samples are randomly drawn from the population.
  - Independent Observations: Observations are independent of each other.

Hypotheses:

- **Two-tailed:**

$$H_0 : \sigma^2 = \sigma_0^2, \quad H_1 : \sigma^2 \neq \sigma_0^2$$

- **Right-tailed:**

$$H_0 : \sigma^2 = \sigma_0^2, \quad H_1 : \sigma^2 > \sigma_0^2$$

- **Left-tailed:**

$$H_0 : \sigma^2 = \sigma_0^2, \quad H_1 : \sigma^2 < \sigma_0^2$$

Test Statistic:

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

s^2 is the sample variance, σ_0^2 is the hypothesized variance and n is the sample size.

Distribution:

$$\chi^2 \sim \chi^2_{(n-1)} \text{ under } H_0$$

Decision Rule:

- Two-tailed: Reject H_0 if $\chi^2 > \chi^2_{\alpha/2,\, n-1}$ or $\chi^2 < \chi^2_{1-\alpha/2,\, n-1}$
- Right-tailed: Reject H_0 if $\chi^2 > \chi^2_{\alpha,\, n-1}$
- Left-tailed: Reject H_0 if $\chi^2 < \chi^2_{1-\alpha,\, n-1}$

α is the significance level and n is the sample size.

Problem:
A biscuit manufacturer claims that the standard deviation of the weight of biscuits in a pack is 2 grams. To verify this claim, a quality control engineer randomly selects 15 biscuits and measures their weights (in grams):

$$50, 49, 52, 51, 50, 48, 53, 49, 51, 50, 52, 48, 49, 51, 50$$

Test, at $\alpha = 0.05$, whether the **variance of biscuit weights** differs from the claimed variance of $\sigma^2 = 4\ \text{g}^2$.

Test for Variance - One Group

Solution

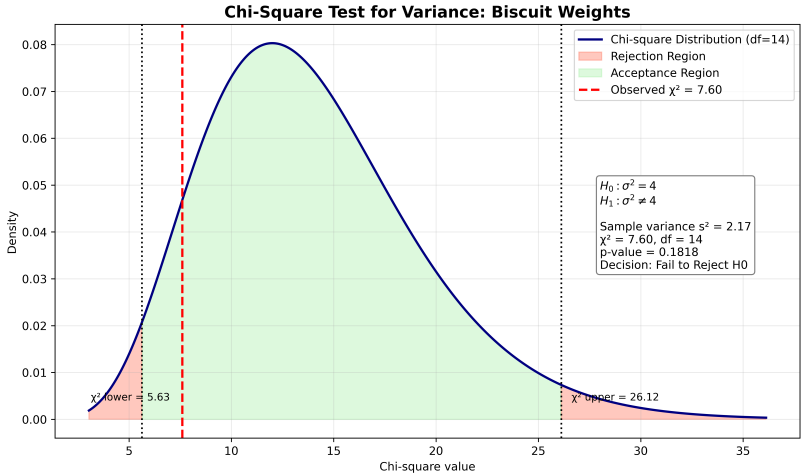


Figure: Chi-Square Test for Variance

Two Group Test

F Test

- The F-test is used to test whether the variances of two populations are equal.
- **Assumptions:**
  - Normality: Both populations are normally distributed.
  - Scale: Data should be measured at interval or ratio scale.
  - Random Sampling: Samples are randomly and independently drawn.
  - Independent Observations: Samples from the two groups are independent of each other.

Hypotheses:

- **Two-tailed:**

$$H_0 : \sigma_1^2 = \sigma_2^2, \quad H_1 : \sigma_1^2 \neq \sigma_2^2$$

- **Right-tailed:**

$$H_0 : \sigma_1^2 = \sigma_2^2, \quad H_1 : \sigma_1^2 > \sigma_2^2$$

- **Left-tailed:**

$$H_0 : \sigma_1^2 = \sigma_2^2, \quad H_1 : \sigma_1^2 < \sigma_2^2$$

Test Statistic:

$$F = \frac{s_1^2}{s_2^2}$$

s_1^2 is larger sample variance and s_2^2 is smaller sample variance

Let

$$X_1, X_2, \ldots, X_{n_1} \sim N(\mu_1, \sigma_1^2), \quad Y_1, Y_2, \ldots, Y_{n_2} \sim N(\mu_2, \sigma_2^2)$$

be two independent samples. We want to test:

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad \text{vs.} \quad H_1 : \sigma_1^2 \neq \sigma_2^2$$

$$s_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_i - \bar{X})^2, \quad s_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2$$

$$\frac{(n_1 - 1)s_1^2}{\sigma_1^2} \sim \chi^2_{n_1 - 1}, \quad \frac{(n_2 - 1)s_2^2}{\sigma_2^2} \sim \chi^2_{n_2 - 1}$$

F-Statistic:

$$F = \frac{\frac{(n_1-1)s_1^2}{\sigma_1^2}/(n_1-1)}{\frac{(n_2-1)s_2^2}{\sigma_2^2}/(n_2-1)} = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$$

If $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma^2$, then

$$F = \frac{s_1^2}{s_2^2} \sim F_{n_1-1,n_2-1}$$

Note: $n_1 - 1$ and $n_2 - 1$ are the numerator and denominator degrees of freedom, respectively.

- Under $H_0 : \sigma_1^2 = \sigma_2^2$, the test statistic follows an F-distribution with $(n_1 - 1, n_2 - 1)$ degrees of freedom:

$$F \sim F_{n_1-1, n_2-1}$$

- **Decision Rule:**
  - Two-tailed: Reject H_0 if $F < F_{1-\alpha/2,\, n_1-1, n_2-1}$ or $F > F_{\alpha/2,\, n_1-1, n_2-1}$
  - Right-tailed: Reject H_0 if $F > F_{\alpha,\, n_1-1, n_2-1}$
  - Left-tailed: Reject H_0 if $F < F_{1-\alpha,\, n_1-1, n_2-1}$

F-Test for Equality of Variances

Problem:
A biscuit company produces biscuits in two different factories, Factory A and Factory B. The quality control team wants to check if the variability in biscuit weights is the same for both factories.
A random sample of biscuit weights (in grams) from each factory is collected:

$$\text{Factory A } (n_1 = 10): 50, 52, 49, 51, 50, 48, 53, 49, 51, 50$$

$$\text{Factory B } (n_2 = 12): 47, 50, 49, 48, 51, 46, 52, 49, 48, 50, 47, 49$$

Test, at $\alpha = 0.05$, whether the **variance of biscuit weights** is the same in the two factories.

F-Test for Equality of Variances

Solution

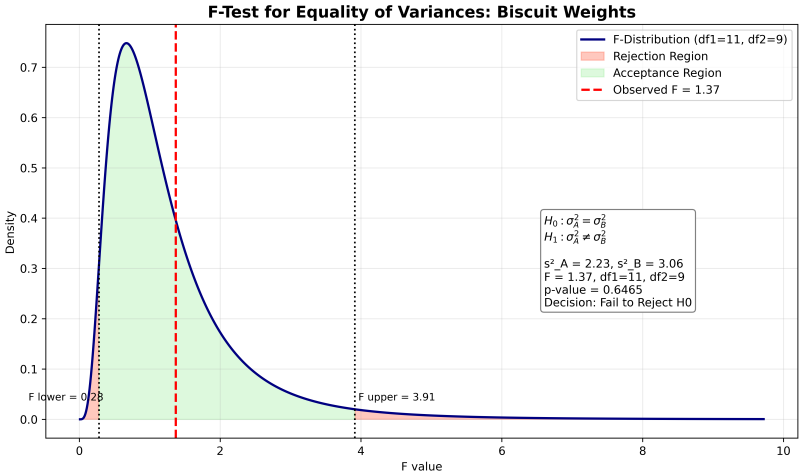


Figure: F-test

Test for Proportion

χ^2 & F Tests

One Group Test

Z, Chi-Square for Independence Test

- Test for Proportion - One Group determine whether the proportion of successes in a population differs significantly from a hypothesized value.
- Assumptions
  - **Random Sampling:** The sample must be randomly selected from the population.
  - **Independent Observations:** Each observation should be independent of the others.
  - **Sample Size Adequacy:** The sample size should be large enough for the normal approximation to be valid:
    $$np_0 \geq 5 \quad \text{and} \quad n(1-p_0) \geq 5$$
    where p_0 is the hypothesized population proportion.
  - **Binary Outcome:** The variable under study should be categorical with two possible outcomes (e.g., success/failure, pass/fail).

- **Two-tailed:**

$$H_0 : p = p_0, \quad H_1 : p \neq p_0$$

- **Right-tailed:**

$$H_0 : p = p_0, \quad H_1 : p > p_0$$

- **Left-tailed:**

$$H_0 : p = p_0, \quad H_1 : p < p_0$$

Test Statistic:

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

where $\hat{p} = \frac{x}{n}$ is the sample proportion and n is the sample size.

Decision Rule:

- Two-tailed: Reject H_0 if $|Z| > Z_{\alpha/2}$
- Right-tailed: Reject H_0 if $Z > Z_{\alpha}$
- Left-tailed: Reject H_0 if $Z < -Z_{\alpha}$

Let $X_1, X_2, \ldots, X_n$ be i.i.d. Bernoulli random variables, where

$$X_i = \begin{cases} 1 & \text{if "success"} \\ 0 & \text{if "failure"} \end{cases}$$

The probability of success is p_0.

Sample proportion:

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

$\hat{p}$ is approximately normally distributed with

$$E[\hat{p}] = p_0, \quad Var(\hat{p}) = \frac{p_0(1-p_0)}{n}, \quad \text{for large } n.$$

Step 1: Mean and variance of $\hat{p}$:

$$E[\hat{p}] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n} \cdot np_0 = p_0$$

$$Var(\hat{p}) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2} \cdot n\, p_0(1-p_0) = \frac{p_0(1-p_0)}{n}$$

Step 2: Apply Central Limit Theorem (CLT):
For large n, the sum of i.i.d. variables $\sum X_i$ is approximately normal:

$$\sum_{i=1}^{n} X_i \sim N(np_0,\ np_0(1-p_0))$$

Dividing by n:

$$\hat{p} = \frac{1}{n}\sum X_i \sim N\Big(p_0,\ \frac{p_0(1-p_0)}{n}\Big)$$

We standardize $\hat{p}$ to create a Z-statistic:

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

$$Z \sim \mathcal{N}(0,1)$$

Step 3: Define significance level and critical region

- Choose a significance level α (commonly 0.05).
- Determine the type of test:
  - **Two-tailed:** $H_0 : p = p_0$ vs $H_1 : p \neq p_0$ Reject H_0 if $|Z| > Z_{\alpha/2}$
  - **Left-tailed:** $H_0 : p = p_0$ vs $H_1 : p < p_0$ Reject H_0 if $Z < -Z_\alpha$
  - **Right-tailed:** $H_0 : p = p_0$ vs $H_1 : p > p_0$ Reject H_0 if $Z > Z_\alpha$

Step 4: Compare calculated Z with critical Z

If Z_{calc} falls in the rejection region, reject H_0

Step 5: Compute p-value (optional)

- p-value $= P(|Z| \geq |Z_{\text{calc}}|)$ for two-tailed
- Reject H_0 if p-value $< \alpha$

Hypotheses for One-Sample Proportion Test

Z-test

A university claims that only 10% of its students use illicit drugs. To verify this claim, a survey is conducted on 200 randomly selected students. It is found that 28 students reported using drugs.
Test, at a significance level of $\alpha = 0.05$, whether the true proportion of students using drugs is different from 0.10.

Hypotheses for One-Sample Proportion Test

Z-test

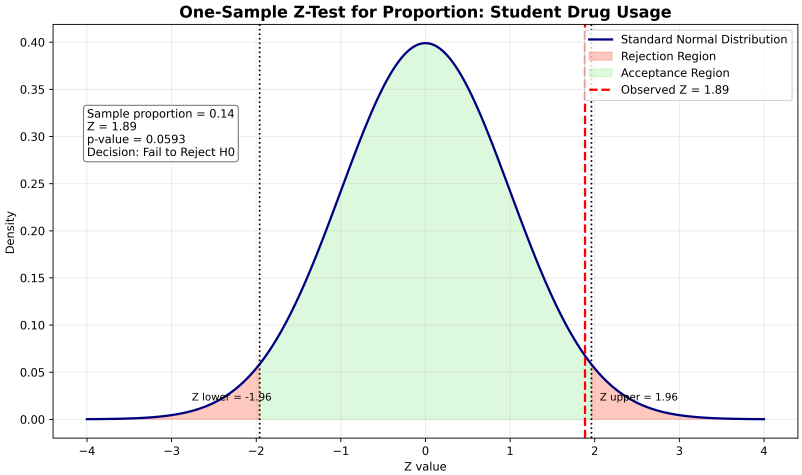


Figure: Z-test for One Group Proportion

Two Group Test

Z-Test for Proportions

- The **two-sample proportion Z-test** determines whether the proportions of successes in two independent populations differ significantly.
- **Assumptions:**
  - **Independent Samples:** The two samples must be drawn independently.
  - **Random Sampling:** Each sample should be randomly selected.
  - **Sample Size Adequacy:** Normal approximation is valid if

$$n_1\hat{p}_1 \geq 5, \quad n_1(1-\hat{p}_1) \geq 5, \quad n_2\hat{p}_2 \geq 5, \quad n_2(1-\hat{p}_2) \geq 5$$

  - **Binary Outcome:** Each observation represents a success or failure.

Hypotheses for Two-Sample Proportion Test

- **Two-tailed Test:**

$$H_0 : p_1 = p_2 \quad \text{vs.} \quad H_1 : p_1 \neq p_2$$

- **Left-tailed Test:**

$$H_0 : p_1 = p_2 \quad \text{vs.} \quad H_1 : p_1 < p_2$$

- **Right-tailed Test:**

$$H_0 : p_1 = p_2 \quad \text{vs.} \quad H_1 : p_1 > p_2$$

Test Statistic and Decision Rule for Two-Sample Proportion Test

- **Test Statistic:**

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where

$$\hat{p}_i = \frac{x_i}{n_i}, \qquad \hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

- **Decision Rule:**
  - **Two-tailed test:** Reject H_0 if $|z| > z_{\alpha/2}$
  - **Right-tailed test:** Reject H_0 if $z > z_\alpha$
  - **Left-tailed test:** Reject H_0 if $z < -z_\alpha$

Multiple Group Test

Chi-Square Test for Independence or Homogeneity

- Test for Proportion - Multiple Groups checks whether two or more categorical variables are statistically independent or whether the distribution of proportions across categories is the same for all groups.
- **Assumptions**
  - **Random Sampling:** The data should be obtained from a random sample of the population.
  - **Independence of Observations:** Each observation must belong to one and only one cell of the contingency table. No individual or item should be counted more than once.
  - **Expected Frequency Condition:** The expected frequency in each cell should be sufficiently large for the chi-square approximation to the sampling distribution to be valid:

$$E_{ij} \geq 5 \quad \text{for most cells.}$$

  - **Measurement Scale:** Both variables should be categorical (nominal or ordinal).

- **Hypotheses:**

$$H_0 : \text{The two categorical variables are independent}$$

$$H_1 : \text{The two categorical variables are not independent}$$

- **Hypotheses:**

$$H_0 : p_1 = p_2 = p_3 = \cdots = p_k$$

$$H_1 : \text{At least one proportion } p_i \text{ differs.}$$

- **Test Statistic:**

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where

$$E_{ij} = \frac{(R_i)(C_j)}{N}$$

O_{ij}: observed frequency,
E_{ij}: expected frequency,
R_i: total of row i,
C_j: total of column j,
N: grand total
r : number of rows
c : number of columns

- **Observed Frequency Matrix:**

$$O = \begin{bmatrix} O_{11} & O_{12} & \dots & O_{1c} \\ O_{21} & O_{22} & \dots & O_{2c} \\ \vdots & \vdots & \ddots & \vdots \\ O_{r1} & O_{r2} & \dots & O_{rc} \end{bmatrix}$$

- **Expected Frequency Matrix:**

$$E = \begin{bmatrix} E_{11} & E_{12} & \dots & E_{1c} \\ E_{21} & E_{22} & \dots & E_{2c} \\ \vdots & \vdots & \ddots & \vdots \\ E_{r1} & E_{r2} & \dots & E_{rc} \end{bmatrix}$$

- The chi-square test with degrees of freedom df = (r-1)(c-1) compares **observed frequencies** O_{ij} with **expected frequencies** E_{ij} under the null hypothesis H_0 of independence:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

  - Under H_0, the counts O_{ij} in each cell can be approximated by a **multinomial distribution**.
  - Each standardized term
$$\frac{O_{ij} - E_{ij}}{\sqrt{E_{ij}}}$$
    is approximately Normal$(0, 1)$ for large sample sizes ($E_{ij} \geq 5$).
  - The sum of the squares of $r \times c$ independent standard normal variables is **Chi-square distributed**.

- Consider a contingency table with r rows and c columns. Each observation falls into exactly one cell (i, j).
- Let N be the total number of observations and p_{ij} the probability of an observation being in cell (i, j). Then the vector of cell counts:

$$(O_{11}, O_{12}, \ldots, O_{rc}) \sim \text{Multinomial}\,(N, (p_{11}, p_{12}, \ldots, p_{rc}))$$

- Under the null hypothesis of independence:

$$\boxed{H_0 : p_{ij} = \frac{R_i C_j}{N^2} \quad \Rightarrow \quad E_{ij} = N p_{ij} = \frac{R_i C_j}{N}}$$

  where R_i and C_j are the row and column totals, respectively.
- The Chi-square test statistic χ^2 compares observed counts O_{ij} to expected counts E_{ij} under the multinomial model.

- Under the null hypothesis H_0 of independence, the chi-square statistic is

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

For large sample size N, this statistic approximately follows a chi-square distribution with $(r-1)(c-1)$ degrees of freedom:

$$\chi^2 \sim \chi^2_{(r-1)(c-1)}.$$

The degrees of freedom arise because:

$$\text{df} = rc - 1 - [(r-1) + (c-1)] = (r-1)(c-1).$$

- **Decision Rule:** Reject H_0 if

$$\chi^2_{\text{calc}} > \chi^2_{\alpha,(r-1)(c-1)}$$

- **Question:** A random sample of 500 U.S. adults was surveyed regarding their political affiliation and opinion on a tax reform bill. Does one's opinion depend on party affiliation?
- **Observed Data (Contingency Table):**

$O =$

Party / Opinion	Favor	Indifferent	Opposed	Total
Democrat	138	83	64	285
Republican	64	67	84	215
Total	202	150	148	500

- **Null Hypothesis (H_0):**

H_0 : Opinion on the tax reform bill is independent of political affiliation.

- **Alternative Hypothesis (H_1):**

H_1 : Opinion on the tax reform bill depends on political affiliation.

Observed counts:

$O =$

Party / Opinion	Favor	Indifferent	Opposed	Row Total
Democrat	138	83	64	285
Republican	64	67	84	215
Column Total	202	150	148	500

Expected counts under H_0:

$E =$

Party / Opinion	Favor	Indifferent	Opposed	Row Total
Democrat	115.14	85.50	84.36	285
Republican	86.86	64.50	63.64	215
Column Total	202	150	148	500

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{3}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} = 22.152, \quad df = 2$$

Critical value: $\chi^2_{0.05,2} = 5.991$

Since $22.152 > 5.991$, we **reject** H_0.

- “If we increase our marketing budget by 1 lakh, how much more can we expect in sales?”
- ”How does employee training investment impact productivity?”
- ”How does advertising expenditure influence sales revenue?”
- ”Both marketing spend and number of sales staff together influence total sales.”
- ”How does product price and household income impact demand”

The linear regression model in the conditional-expectation form is defined as:

$$\boxed{\mathbb{E}[Y \mid X] = \beta_0 + \beta^\top X}$$

- Y is the dependent variable
- X is the set of independent variables $X_1, X_2, \ldots X_p$
- p is the number of independent variables
- β is the set of p+1 parameters $\beta_o, \beta_1, \ldots, \beta_p$
- $\mathbb{E}[Y \mid X]$ is the prediction made by the model given the set of p independent variables $X_1, X_2, \ldots X_p$

$$\boxed{\mathbb{E}[Y \mid X] = \beta_0.1 + \beta_1.X_1 + \beta_2.X_2 \ldots \beta_p.X_p}$$

$$Y = \mathbb{E}[Y \mid X] + \varepsilon$$

$$\boxed{Y = \beta_0 + \beta^\top X + \varepsilon, \qquad \text{where } \mathbb{E}[\varepsilon \mid X] = 0.}$$

$$(\beta_0^*, \beta^*) = \arg \min_{\beta_0 \in \mathbb{R},\, \beta \in \mathbb{R}^p} \mathbb{E}\big[(Y - \beta_0 + \beta^\top X)^2\big].$$

The minimizer (β_0^*, β^*) satisfies the population normal equations:

$$\boxed{\mathbb{E}[Y] = \beta_0^* + (\beta^*)^\top \mathbb{E}[X], \qquad \mathbb{E}\big[X(Y - \beta_0^* - (\beta^*)^\top X)\big] = 0.}$$

- The expected (average) value of Y in the population equals the predicted value from the regression line, when evaluated at the average of X.
- The "unexplained part" of Y or the error $\epsilon = Y - \beta_0^* - (\beta^*)^\top X$ is uncorrelated with all the predictors X.

- Given data $\{(y_i, x_{i1}, x_{i2}.\ldots.x_{ip})\}_{i=1}^{n}$, the OLS estimator solves:

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{n}\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1.x_{1i} + \cdots + \beta_p.x_{pi}))^2.$$

-

$$\boxed{\hat{\beta}_{aug} = (X_{\text{aug}}^{\top}X_{\text{aug}})^{-1}X_{\text{aug}}^{\top}y}$$

$$\boxed{\mathbb{E}[Y \mid X] = \hat{\beta}_{aug}^{\top}X}$$

This is the equation of the line or plane that predicts Y from X by finding the coefficients that minimize the overall squared errors between the predicted and actual values.

Linear Regression

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 . x_{1i} + \cdots + \beta_p . x_{pi}))^2.$$

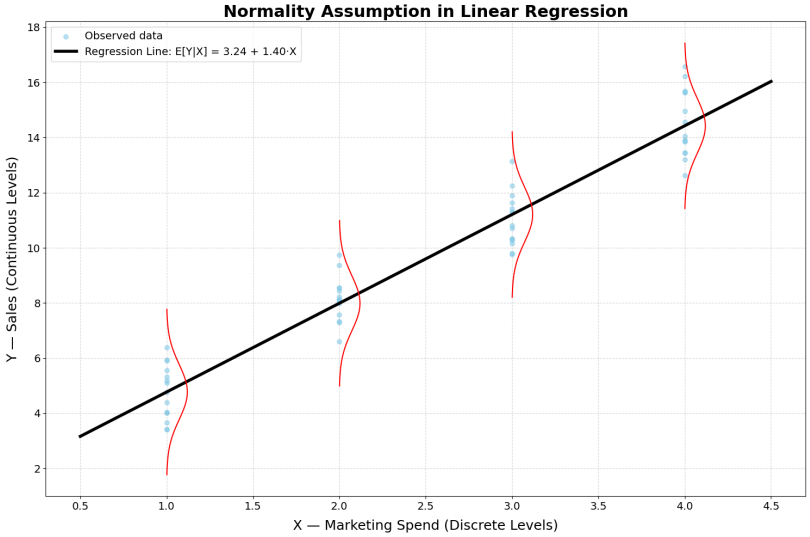
$$\boxed{\hat{\beta} = (X_{\text{aug}}^\top X_{\text{aug}})^{-1} X_{\text{aug}}^\top y}$$

$$X_{\text{aug}} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}_{n \times (p+1)}$$

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{n \times 1}$$

Linear Regression

Normality Assumption



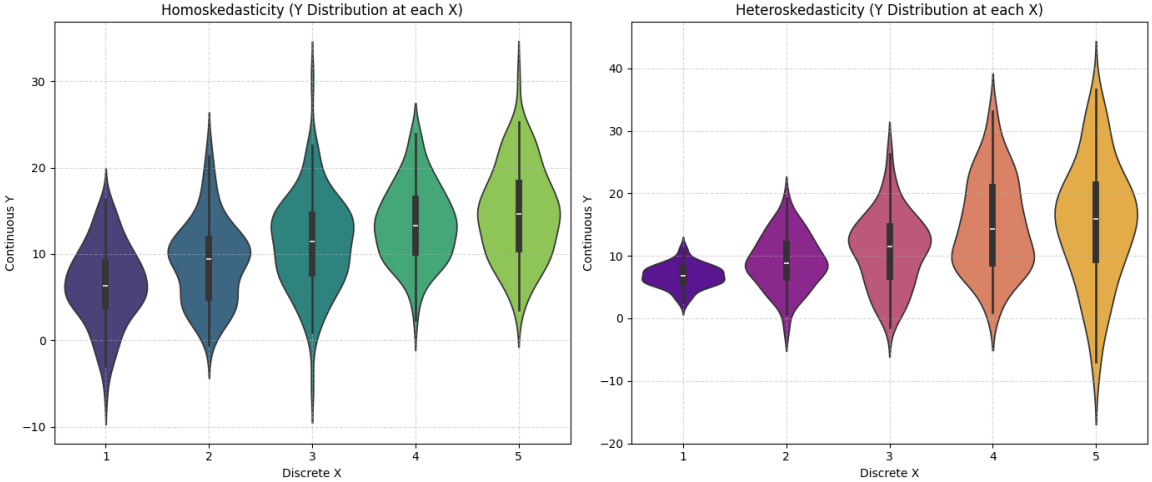


Figure: Linear Regression - Homo-skedasticity Assumption

Sample Data (15 observations):

Obs	Marketing Spend (k)	Sales Staff	Sales Revenue (k)
1	10	2	25
2	12	3	30
3	15	4	38
4	8	2	20
5	20	5	50
6	18	4	45
7	25	6	60
8	5	1	12
9	14	3	35
10	22	5	55
11	9	2	22
12	16	4	40
13	13	3	33
14	7	1	15
15	19	5	48

Predictor	**Estimate**	**Std. Error**	**t value**	**Pr($>\|t\|$)**
(Intercept β_0)	5.00	1.50	3.33	0.005
Marketing Spend β_1	1.50	0.10	15.00	< 0.001
Sales Staff β_2	3.00	0.20	15.00	< 0.001

Table: OLS Regression Results

R-squared = 0.98, **Adjusted R-squared** = 0.97

$$\text{Sales Revenue} = \hat{\beta}_0 + \hat{\beta}_1 \times \text{Marketing Spend} + \hat{\beta}_2 \times \text{Sales Staff}$$

The correlation coefficient between independent variables should be minimum.

	Marketing Spend	**Sales Staff**	**Sales Revenue**
Marketing Spend	1.00	0.98	0.99
Sales Staff	0.98	1.00	0.99
Sales Revenue	0.99	0.99	1.00

Hence, ideally one of the variables need to be removed.

Fitted Model:

$$\text{Sales Revenue} = \hat{\beta}_0 + \hat{\beta}_1 \times \text{Marketing Spend}$$

Predictor	**Estimate**	**Std. Error**	**t value**	**Pr($>|t|$)**
(Intercept)	10.00	1.20	8.33	0.0001
Marketing Spend	2.00	0.12	16.67	< 0.001

Table: OLS Regression Results (Simple Linear Model)

R-squared = 0.95, **Adjusted R-squared** = 0.95

The Pearson correlation coefficient r measures the strength and direction of the *linear relationship* between two continuous variables X and Y.

$$r_{XY} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

Interpretation:

- $r = 1$: Perfect positive linear relationship
- $r = -1$: Perfect negative linear relationship
- $r = 0$: No linear relationship

Pearson Correlation Coefficient:

$$\rho_{XY} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\mathbb{E}\big[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\big]}{\sqrt{\mathbb{E}\big[(X - \mathbb{E}[X])^2\big]\, \mathbb{E}\big[(Y - \mathbb{E}[Y])^2\big]}}$$

- $\mathbb{E}[X]$ = expected value (mean) of X
- $\sigma_X = \sqrt{\mathbb{E}[(X - \mathbb{E}[X])^2]}$ = standard deviation of X
- $\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$

Covariance vs. Correlation:

Property	**Covariance** $\mathrm{Cov}(X, Y)$	**Correlation** ρ_{XY}
Definition	$\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$	$\frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}$
Unit dependence	Depends on units of X and Y	Unit-free (dimensionless)
Range	Unbounded	$[-1, +1]$
Interpretation	Direction of linear relationship	Strength and direction of linear relationship
Scale sensitivity	Changes with variable scaling	Invariant to scaling (standardized)

Correlation of Linearly Scaled Variables

$$Corr(aX, bY) = sign(ab)\text{Corr}(X, Y) \tag{20}$$

$$Cov(aX, bY) = (ab)\text{Cov}(X, Y) \tag{21}$$

- Correlation is scale-invariant — only the sign of a and b matters.
- If $a, b > 0$, correlation remains unchanged.
- If one of a or b is negative, correlation changes sign.
- Covariance is scale variant. Covariance magnitude depends upon the product of a and b.

Let X and Y be random variables with means μ_X, μ_Y and standard deviations σ_X, σ_Y. The normalized (standardized) forms are:

$$Z_X = \frac{X - \mu_X}{\sigma_X}, \qquad Z_Y = \frac{Y - \mu_Y}{\sigma_Y}$$

Then:

$$E[Z_X] = E[Z_Y] = 0, \qquad \text{Var}(Z_X) = \text{Var}(Z_Y) = 1$$

$$\begin{aligned}
\text{Cov}(Z_X, Z_Y) &= E[(Z_X - E[Z_X])(Z_Y - E[Z_Y])] \\
&= E[Z_X Z_Y] \\
&= E\left[\frac{(X - \mu_X)(Y - \mu_Y)}{\sigma_X \sigma_Y}\right] \\
&= \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \rho(X, Y)
\end{aligned}$$

Spearman Correlation measures the *monotonic relationship* between two variables using their *ranks*. Given n paired observations (x_i, y_i):

1. Convert each variable to ranks:
$$R(x_i),\ R(y_i)$$
2. Compute the difference in ranks:
$$d_i = R(x_i) - R(y_i)$$
3. The Spearman coefficient is given by:
$$r_s = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

r_s Value	Interpretation
$+1$	Perfect positive monotonic relationship
0	No monotonic relationship
-1	Perfect negative monotonic relationship

Aspect	Pearson Correlation	Spearman Correlation
Relationship	Linear relationship	Monotonic relationship (increasing or decreasing)
Data Type	Requires interval or ratio scale	Works Ordinal, interval, or ratio data
Assumption	Assumes normality	Non-parametric
Effect of Outliers	Highly sensitive	Less sensitive

Table: Pearson and Spearman Correlation Coefficient

Correlation Coefficient

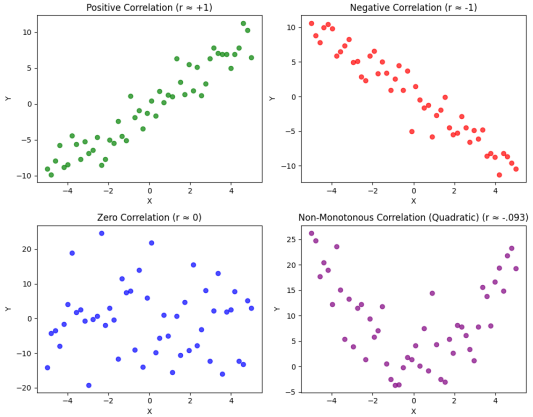
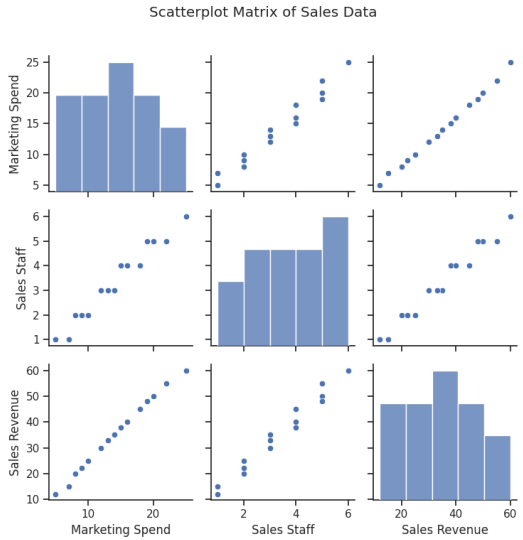


Figure: Correlation Coefficient

Linear Regression

Correlation from Scatter Plot



Linear Regression

Scatter Matrix and Correlation Matrix

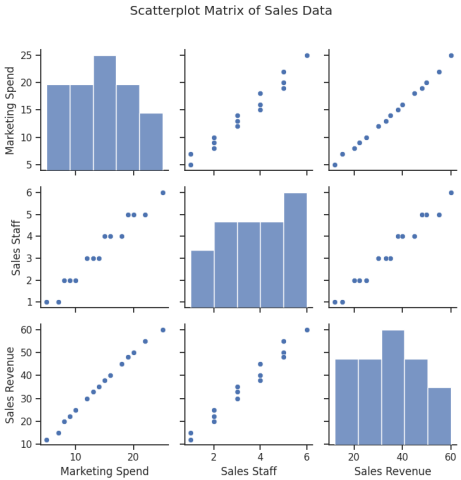


Figure: Scatter Matrix

	MS	SS	SR
Marketing Spend MS	1.00	0.98	0.99
Sales Staff SS	0.98	1.00	0.99
Sales Revenue SR	0.99	0.99	1.00

Definition:

$$R^2 = 1 - \frac{\text{SSR}}{\text{SST}}$$

where

$$\text{SSR} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \qquad \text{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

- y_i: Observed values
- $\hat{y}_i$: Predicted values from the regression model
- $\bar{y}$: Mean of observed y_i
- SSR (Sum of Squared Residuals): Unexplained variation
- SST (Total Sum of Squares): Total variation in Y

Equivalent Form (Correlation-based):

$$\boxed{R^2 = \left(\text{Corr}(Y, \hat{Y}) \right)^2}$$

Definition:

$$R^2 = 1 - \frac{\text{SSR}}{\text{SST}} = \frac{\text{SST} - \text{SSR}}{\text{SST}} = \frac{\text{Explained Variance}}{\text{Total Variance}}$$

As a Percentage:

$$\text{Percentage of Variance Explained} = R^2 \times 100\%$$

Interpretation:

- R^2 represents the proportion of total variation in Y that is explained by the regression model.
- Expressing it as $R^2 \times 100\%$ gives the percentage of variance in Y explained by the model.
- For example, $R^2 = 0.85$ means the model explains 85% of the variation in Y.

Objective : Model Sales Revenues as a function of Advertising Spend AS, Promotions spend PS, Training hours TH

$$\boxed{\mathbb{E}[Y \mid X] = \beta_0 + \beta_1.AS + \beta_2.PS + \beta_3.TH}$$

Correlation Matrix

	AS	PS	TH
Advertising Spend	1.000000	0.927709	-0.028591
Promotions Spend	0.927709	1.000000	0.005904
Training Hours	-0.028591	0.005904	1.000000

Dep. Variable:	Sales_Revenue	**R-squared:**	0.652
Model:	OLS	**Adj. R-squared:**	0.641
Method:	Least Squares	**F-statistic:**	59.91
Date:	Mon, 27 Oct 2025	**Prob (F-statistic):**	6.47e-22
Time:	02:58:46	**Log-Likelihood:**	-428.41
No. Observations:	100	**AIC:**	864.8
Df Residuals:	96	**BIC:**	875.2
Df Model:	3		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P$>$ |t|**	**[0.025**	**0.975]**
const	17.9809	4.776	3.765	0.000	8.501	27.460
Advertising_Spend A	0.7550	0.163	4.638	0.000	0.432	1.078
Promotions_Spend P	-0.0500	0.200	-0.250	0.803	-0.447	0.347
Training_Hours T	0.8299	0.126	6.603	0.000	0.580	1.079

Omnibus:	1.663	**Durbin-Watson:**	2.324
Prob(Omnibus):	0.435	**Jarque-Bera (JB):**	1.110
Skew:	-0.151	**Prob(JB):**	0.574
Kurtosis:	3.419	**Cond. No.**	199.

Table: Linear Regression Estimator ($X = X_{aug}$)

Quantity	Mathematical Expression
Estimated Coefficient ($\hat{\beta}_j$)	$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$
Variance of Coefficient	$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$
Estimated Variance	$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}, \quad \hat{\sigma}^2 = \frac{\sum(y_i - \hat{y}_i)^2}{n-k-1}$
Standard Error of Coefficient	$SE(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2\,[(\mathbf{X}'\mathbf{X})^{-1}]_{jj}}$
t-statistic	$t_j = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$

Model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(0, \sigma^2 I)$$

OLS Estimator

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

Variance-Covariance Matrix

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}$$

Estimated as:

$$\hat{\sigma}^2 = \frac{\text{SSE}}{n - k}$$

Standard Error of Coefficient

$$SE(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 \left[(\mathbf{X}^\top \mathbf{X})^{-1}\right]_{jj}}$$

t-statistic

$$t_j = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \quad \text{with } (n-k-1) \text{ d.f.}$$

Model

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

Hypotheses for a Coefficient β_j

$H_0 : \beta_j = 0$ (no effect)

$H_1 : \beta_j \neq 0$ (has effect)

t-statistic

$$t_j = \frac{\hat{\beta}_j - 0}{SE(\hat{\beta}_j)}$$

$t_j \sim t_{(n-k-1)}$ under H_0

Decision Rule

Reject H_0 if:

$$|t_j| > t_{\alpha/2,(n-k-1)} \quad \text{or} \quad p\text{-value} < \alpha$$

Interpretation

- Reject H_0: variable x_j significantly affects y.
- Fail to reject H_0: no significant effect of x_j.

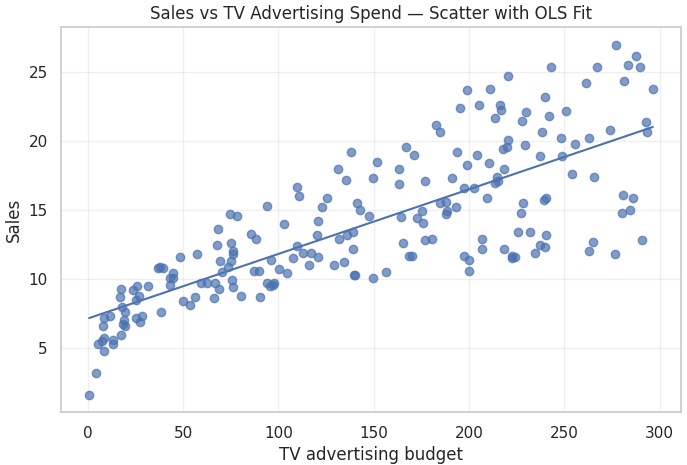


Figure: Sales y $\sim$ TV Advertising Budget x_1 : Data (200*2)

Model Summary and Diagnostics

Table: Summary of Linear Regression Model

Dataset Information	
Number of Observations	200
Number of Variables	4 (TV, Radio, Newspaper, Sales)
Sample Rows	
1	TV = 230.1, Radio = 37.8, Newspaper = 69.2, Sales = 22.1
2	TV = 44.5, Radio = 39.3, Newspaper = 45.1, Sales = 10.4
3	TV = 17.2, Radio = 45.9, Newspaper = 69.3, Sales = 9.3
Fitted Model	$\hat{y} = 7.1318 + 0.0469x$
Model Performance	
R^2_{Train}	0.5907
$\text{MSE}_{\text{Train}}$	10.9612
R^2_{Test}	0.6606
MSE_{Test}	9.1793
Residual Diagnostics	
Durbin–Watson Statistic	1.9985
Shapiro–Wilk (Test Residuals)	$W = 0.9635$, $p = 0.1240$

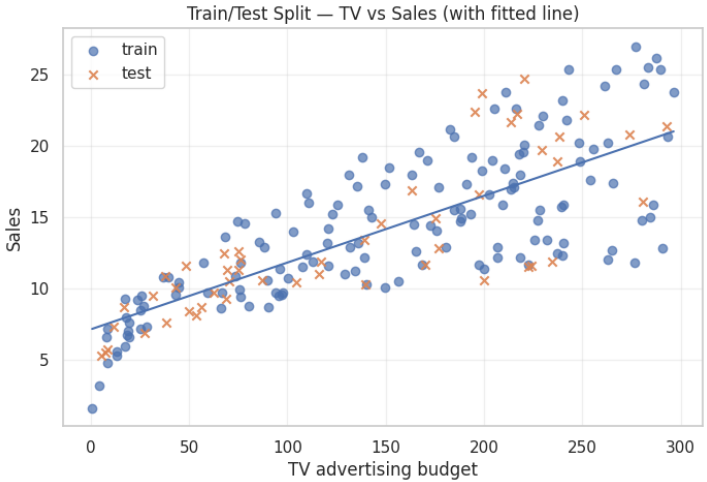


Figure: Training Data : Testing Data - 75:25

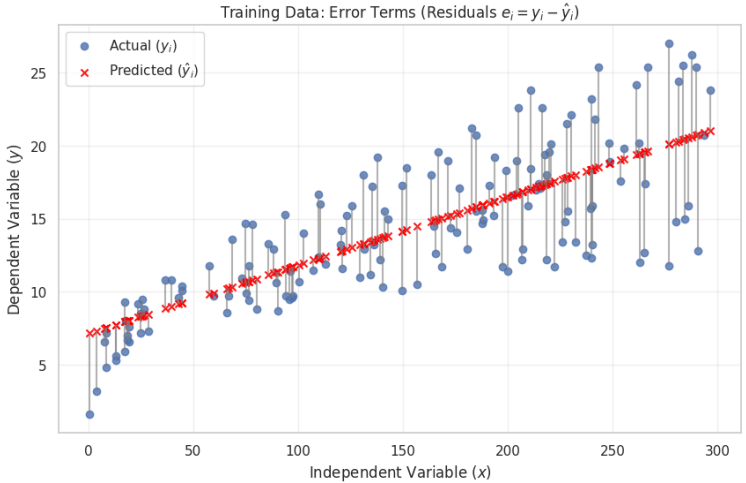


Figure: Error in Linear Regression Model

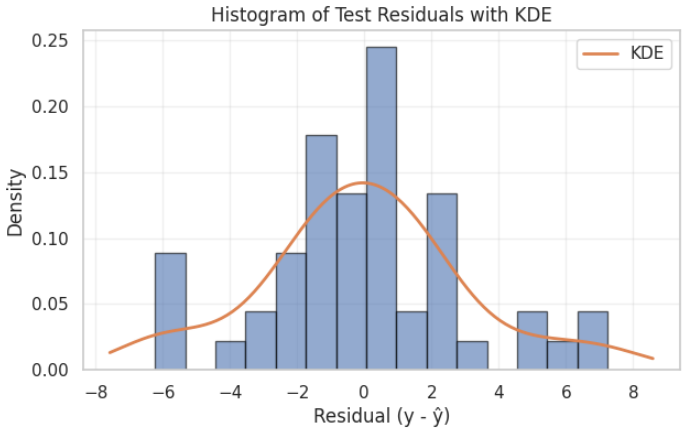


Figure: Residual Analysis

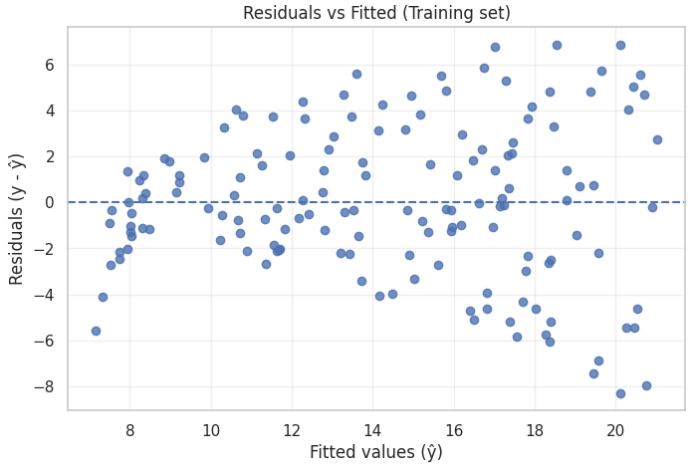


Figure: Residual Analysis

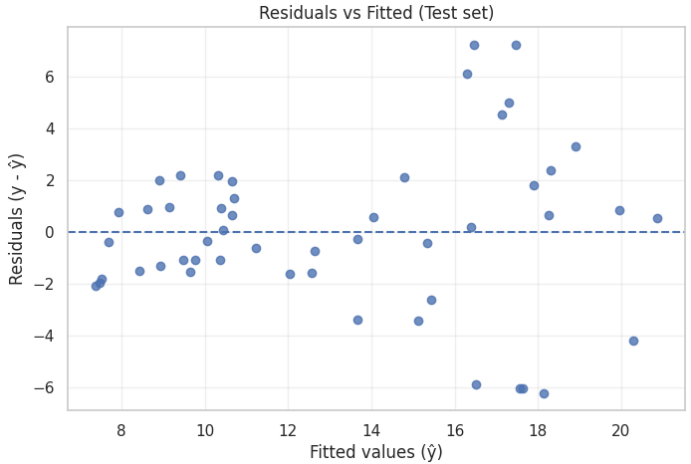
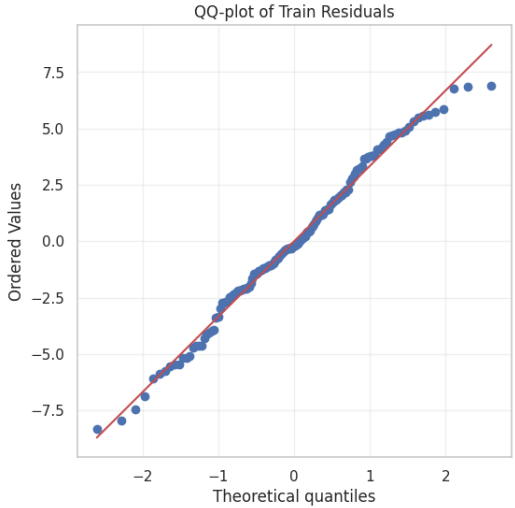
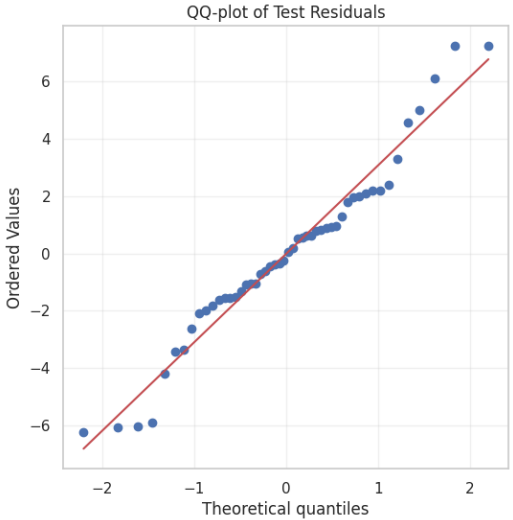


Figure: Residual Analysis


QQ-plot of Train Residuals
Ordered Values
Theoretical quantiles
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−2
−1
0
1
2

Simple Linear Regression - QQ Plot Testing Set



Model Equation:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

Slope (OLS Estimator):

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\mathrm{Cov}(x, y)}{\mathrm{Var}(x)} = r\,\frac{s_y}{s_x}$$

where r is the Pearson correlation coefficient, and s_x, s_y are the sample standard deviations of x and y.

Intercept:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Coefficient of Determination:

$$R^2 = \frac{\text{SS}_{\text{reg}}}{\text{SS}_{\text{tot}}} = 1 - \frac{\text{SS}_{\text{res}}}{\text{SS}_{\text{tot}}} = r^2$$

$$\text{where } \text{SS}_{\text{tot}} = \sum_i (y_i - \bar{y})^2, \quad \text{SS}_{\text{res}} = \sum_i (y_i - \hat{y}_i)^2, \quad \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Mean Squared Error (MSE):

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Unbiased Estimator of Error Variance:

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Key Identities (for OLS with intercept):

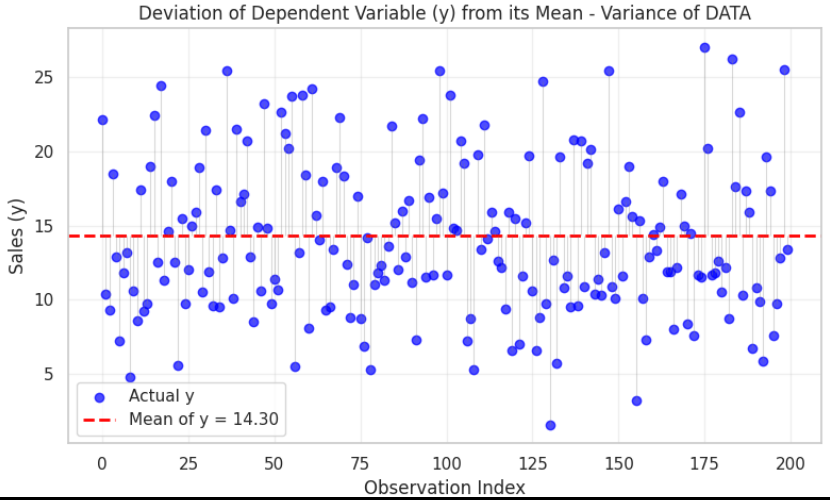
$$\sum_{i=1}^{n}(y_i - \hat{y}_i) = 0, \quad \sum_{i=1}^{n}\hat{y}_i(y_i - \hat{y}_i) = 0 \quad \sum_{i=1}^{n}x_i(y_i - \hat{y}_i) = 0$$

No.	Expression	Interpretation
1	$\sum_{i=1}^{n}(y_i - \hat{y}_i) = 0$	The residuals sum to zero, ensuring that the regression line passes through the mean point $(\bar{x}, \bar{y})$.
2	$\sum_{i=1}^{n}x_i(y_i - \hat{y}_i) = 0$	The residuals are uncorrelated (orthogonal) with the independent variable x. This guarantees the best linear unbiased estimate (BLUE) under Gauss–Markov assumptions.
3	$\sum_{i=1}^{n}\hat{y}_i(y_i - \hat{y}_i) = 0$	The residuals are also uncorrelated with the fitted values $\hat{y}_i$. The explained $\hat{y}_i$ and unexplained components $y_i - \hat{y}_i$ of y are orthogonal.

Linear Regression

SS_{tot} - Measure of Variance of Dependent Variable (Training Set)

$\text{SS}_{\text{tot}} = \sum_i (y_i - \bar{y})^2 = 4017.046$



$$\text{Regression Sum of Squares (SSR)} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$$\text{Total Sum of Squares (SST)} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

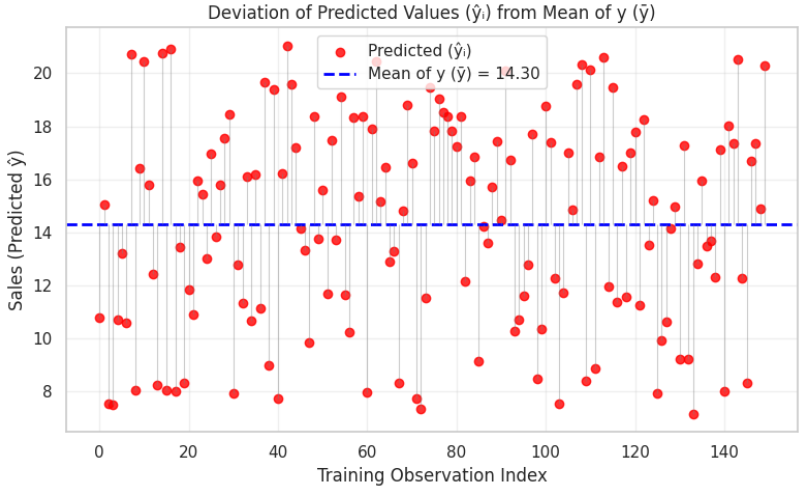
$$\text{Error Sum of Squares (SSE)} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

SST = SSR + SSE

Linear Regression

SS_{tot} - Measure of Variance of Model Predicted Dependent Variable (Training Set)

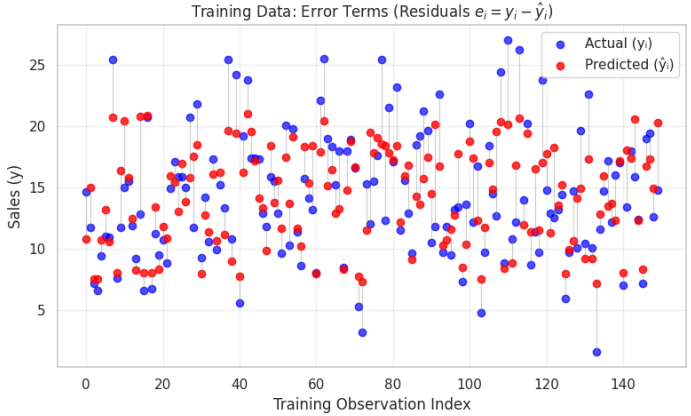
$\text{SS}_{\text{reg}} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = 2372.8716$



Linear Regression

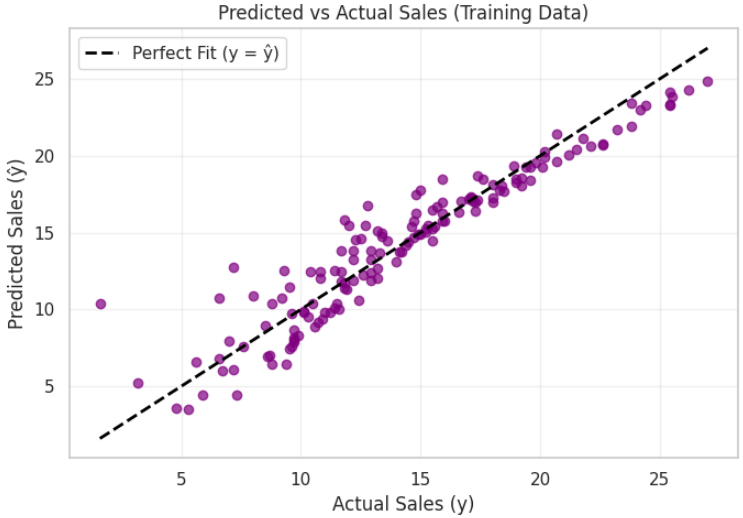
SS$_{error}$ - Measure of Variance of Model Error (Training Set)

$$R^2 = \frac{SS_{reg}}{SS_{tot}} = \frac{2372.8716}{4017.0467} = .5907$$



Multiple Linear Regression $R^2 = \rho^2$

Sales $\sim$ TV, Radio, Newspaper



ANOVA

Analysis of Variance (ANOVA) is used when:

- The independent variable (X) is **categorical** with k groups.
- The dependent variable (Y) is **continuous**.

ANOVA Model

For a one-way ANOVA:

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2)$$

where:

- μ is the overall mean,
- τ_i is the effect of the i^{th} group,
- ε_{ij} is the random error for observation j in group i.

Hypothesis Testing

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k \quad \text{(all group means equal)}$$

$$H_A : \text{At least one mean differs}$$

Test Statistic

$$F = \frac{\text{Between-Group Variance } (\text{MS}_B)}{\text{Within-Group Variance } (\text{MS}_W)} = \frac{\text{SS}_B/(k-1)}{\text{SS}_W/(N-k)}$$

If $F > F_{\alpha,k-1,N-k}$, reject H_0.

- ANOVA generalizes the two-sample t-test to more than two groups.
- A significant F indicates that at least one group mean differs.

- Independence - Observations are independent both within and between groups.
- Normality - The residuals (errors) in each group are normally distributed.
- Homogeneity of variances - The variances across the groups are equal.
- ANOVA assumes each group's data are drawn from a normally distributed population with population mean μ_i, where $\mu_i = \mu + \tau_i$, μ is the overall mean, τ_i is the effect of group i.
- ANOVA is robust to modest violations of normality, especially when group sizes are equal. Hence, we assume group size as equal.

A public health researcher is studying whether three types of physical therapy differ in their effectiveness at improving **post-surgery recovery speed** (measured as the number of **days needed for a patient to regain normal mobility**).
Three types of therapy are randomly assigned to $n = 30$ knee-surgery patients (10 per group):

Group	Therapy Type	Description
1	Conventional Therapy (CT)	Standard hospital-guided physiotherapy.
2	Hydrotherapy (HT)	Exercises performed in a temperature-controlled pool.
3	Robot-Assisted Therapy (RT)	Sessions using robotic gait-assist devices.

The dependent variable is **Recovery Days** (continuous variable). Lower recovery days imply faster recovery and hence a more effective treatment.

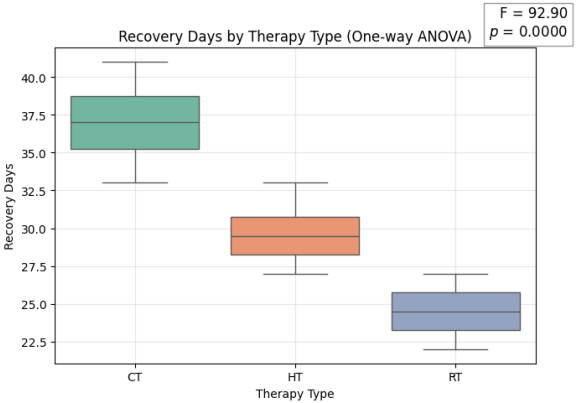
Patient Group	Sample Recovery Days (10 patients each)
CT	33, 37, 41, 35, 39, 36, 38, 34, 40, 37
HT	28, 30, 27, 31, 29, 33, 30, 32, 28, 29
RT	24, 25, 23, 26, 27, 22, 25, 24, 26, 23

Is there a statistically significant difference in the mean recovery time among the three therapy methods?

- **Independent Variable:** Therapy Type (categorical, 3 levels: CT, HT, RT)
- **Dependent Variable:** Recovery Days (continuous)
- **Sample Size:** $n = 30$ (10 per group)

$H_0 : \mu_{CT} = \mu_{HT} = \mu_{RT}$

H_1 : At least one group mean is different.



1. Total Degrees of Freedom:

$$\mathrm{df}_{\mathrm{Total}} = n - 1$$

Measures how all observations vary around the grand mean.

2. Between-Group Degrees of Freedom:

$$\mathrm{df}_{\mathrm{Between}} = k - 1$$

There are k group means, but only $k - 1$ can vary independently.

3. Within-Group Degrees of Freedom:

$$\mathrm{df}_{\mathrm{Within}} = \sum_{i=1}^{k} (n_i - 1) = n - k$$

Each group contributes $(n_i - 1)$ degrees of freedom around its own mean.

Data summary: $n = 30$ observations, $k = 3$ groups (CT, HT, RT).
Grand mean: $\bar{y} = 30.4$.

Source	SS	df	MS = SS/df	F = MS_B/MS_W	p-value
Between Groups (Regression)	788.6000	2	394.3000	92.8979	¡0.001
Within Groups (Error / Residual)	114.6000	27	4.2444		
Total	903.2000	29	—		

$$SS_{Total} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = 903.2,$$

$$SS_{Between} = \sum_{g=1}^{k} n_g (\bar{y}_g - \bar{y})^2 = 788.6,$$

$$SS_{Within} = \sum_{g=1}^{k} \sum_{i \in g} (y_i - \bar{y}_g)^2 = 114.6.$$

$$\mathrm{df}_{\text{Between}} = k - 1 = 2, \quad \mathrm{df}_{\text{Within}} = n - k = 27,$$

$$\mathrm{MS}_{\text{Between}} = \frac{788.6}{2} = 394.3000$$

$$\mathrm{MS}_{\text{Within}} = \frac{114.6}{27} \approx 4.2444$$

$$F = \frac{\mathrm{MS}_{\text{Between}}}{\mathrm{MS}_{\text{Within}}} = \frac{394.3000}{4.2444} \approx 92.8979$$

Conclusion: The observed $F \approx 92.90$ with $\mathrm{df} = (2, 27)$ gives a p-value $\ll 0.001$. Thus we reject H_0 and conclude at usual significance levels that *not all group means are equal* (i.e., at least one therapy differs in mean recovery days).

After finding a significant F-statistic in ANOVA, Tukey's Honest Significance Test (HSD) is used to determine *which specific group means* differ significantly, while **controlling the familywise error rate (FWER)**.

Test statistic:

$$q = \frac{|\bar{Y}_i - \bar{Y}_j|}{\sqrt{\frac{MS_{Within}}{n}}}$$

where:

- $\bar{Y}_i, \bar{Y}_j$ = sample means of groups i and j
- MSE = Mean Square Error (pooled within-group variance from ANOVA)
- n = number of observations per group (assumed equal)

Decision rule:

$$\text{Reject } H_0 \text{ if } q > q_{\alpha, k, df_{\text{within}}}$$

where $q_{\alpha,k,df_{\text{within}}}$ is the critical value from the *studentized range distribution*, α is the level of significance, k is the number of groups and df_{within} or n-k is the degreees of freedom of within variance.

Interpretation:

- If $q > q_{\text{critical}}$, the two group means differ significantly.
- Tukey's HSD adjusts for multiple comparisons, keeping the overall α at the desired level (e.g., 0.05).

Why not run multiple T-Tests instead of Tukey's HSD?

- For k groups, m= $\binom{K}{2}$ combinations of T-Tests has to be conducted. For eg: Groups A,B,C - T-Tests to be conducted for Groups A & B,B & C,C & A.
- The family wise error rate FWER is the probability that at least one of these tests incorrectly rejects a true null hypothesis or the probability of making atleast one type 1 error.
- FWER = 1 - $(1-\alpha)^m$, where m is the total number of T-tests to be conducted. For $\alpha = .05$ and m = 10, FWER = .401 or there is at-least 40% of chance of making one rejection of null hypotheses when it is indeed true.
- Effectively, using T-test instead of Tukey's HSD might result in rejection of the null hypothesis, when it is indeed true.