Prof. (Dr.) R.M. KAPILA RATHNAYAKA

FAHAD

Statistics

for Experiment Analysis

SABARAGAMUWA UNIVERSITY OF SRI LANKA

- Estimation
- Hypothesis Testing
- Introduction to Design of Experiments
- Two Way ANOVA, Special experiment designs
- Design and Analysis of an Experiment
- Chi-Squared test of goodness of fit
- Real-World data analysis with statistical Software

PST 21209

TABLE OF CONTENTS

BASIC DEFINITIONS AND CONCEPTS ... 1

Kinds of Statistics ... 1

KEY STATISTICAL CONCEPTS ... 1

Population ... 1

Mainly the term population can be divided in to two parts. ... 1

Sample ... 2

Reasons for sampling ... 2

Collecting data from a sample ... 2

Fundamentals methods for selecting sample ... 2

Acceptance Sampling ... 3

Sampling Techniques ... 4

Probability Sampling ... 4

Simple Random Sampling ... 4

Stratified Sampling ... 5

Systematic Sampling ... 6

Cluster Sampling ... 6

Statistical Inference ... 7

STATISTICAL ESTIMATION ... 8

POINT ESTIMATION ... 8

Properties of Point Estimators ... 8

Bias ... 8

Unbiased Estimates ... 8

Bias of an Estimate ... 9

Consistency ... 9

CONFIDENCE INTERVALS ... 10

What Are Confidence Intervals? ... 10

Why do we need Confidence Intervals? ... 10

Interval Estimation ... 10

How to Estimate Confidence Intervals for Sample Mean ... 10

Conditions Needed: ... 10

Interval Estimation for population mean μ when $\sigma 2$ is known ... 11

Interval Estimation for Population mean μ when $\sigma 2$ is unknown and n $\leq$ 30 ... 13

Interval Estimation for Population mean μ when $\sigma 2$ is unknown and n $>$ 30 ... 15

Interval Estimation for two samples ... 15

Estimation Requirements ... 15

Two Samples: Estimating the Difference between Two Means ... 15

Interval Estimation of difference of the means of Two Populations when $\sigma 1, \sigma 2$ are Unknown and $n1, n2 \leq 30$ ... 16

CONFIDENCE INTERVAL FOR A POPULATION PROPORTION ... 18

Estimation Requirements ... 18

Single Sample: Estimating a Proportion ... 18

Compare the two proportions by finding Difference in Popular Proportion ... 20

Single Sample: Estimating the Variance ... 21

HYPOTHESIS TESTING ... 23

What is Statistical Hypothesis Testing? ... 23

Why is it Important? ... 23

Two Main Types of Statistical Inference: ... 23

Examples of Statistical Hypothesis Testing: ... 23

What is a Statistical Hypothesis? ... 23

Null Hypothesis (H_0): ... 23

Alternative Hypothesis (H_1 or H_a): ... 24

HYPOTHESIS TESTS ... 24

There are five ingredients to any statistical test ... 25

Test Statistics ... 26

Rejection/Critical Region ... 26

Conclusions ... 27

HYPOTHESIS TEST: DIFFERENCE BETWEEN MEANS 30
COMPARING TWO POPULATION MEANS: PAIRED DATA THE PAIRED T- PROCEDURE 32
Hypothesis Conditions *32*
Test Statistic *32*
ANOVA (ANALYSIS OF VARIANCE) 34
WHAT IS ANOVA? 34
WHY USE ANOVA (INSTEAD OF MANY T-TESTS)? 34
BASIC IDEA 34
WHAT IT DOES 34
LAYOUT OF THE DESIGN 34
ONE-WAY ANOVA 35
What is this test for? *35*
Models for the Data *35*
Means Model 35
Effect Model 35
Assumptions and Hypothesis in One-way ANOVA *36*
Assumptions 36
Hypothesis 36
ANOVA Table for CRD of a Single Factor Experiment: One Way ANOVA Table *37*
Total Sum Square (SS_T) 37
Total Variance due to treatment (SS_{Tr}) 37
TWO-WAY ANOVA 39
Assumptions 40
Hypotheses 41
Total Sum Square (SS_T) 41
Total Variance due to treatment (SS_{Tr}) 41
Total Variance due to block (SS_B) 42
ANALYSIS OF VARIANCE OF A RANDOMIZED BLOCK DESIGN WITH MISSING OBSERVATIONS 44
THE LATIN SQUARE DESIGN 45
SAMPLE LAYOUT: 45
DEFINITION 45
STRUCTURE 45
Parametric model *46*
Constrains *46*
Assumption *46*
Hypothesis *47*
Total Sum Square (SS_T) 47
Total Variance due to treatment (SS_{Tr}) 47
Total Variance due to row (SS_R) 48
Total Variance due to Column (SS_C) 48
LATIN SQUARE DESIGN WITH MISSING OBSERVATIONS 50

EQUATIONS

Equation 1: Formula for determining the sample size 3
Equation 2: Equation for determine confidence Interval 11
Equation 3: Equation for determine confidence Interval when we don't know σ 13
Equation 4: Equation for determine confidence Interval when we don’t know Σ and n value too large as well 15
Equation 5: Estimating the Difference between Two Means 15
Equation 6: pooled estimate of the population standard deviation 16
Equation 7:Equation for determine confidence Interval Difference when we don't know Σ 17
Equation 8: Confidence Interval for Population Proportion 19
Equation 9: C.I. for the difference in two population proportions 20
Equation 10: Confidence Interval for Variance & Standard Deviation of population 21
Equation 11: Test statistics for Paired data 32
Equation 12: Treatment Total 35
Equation 13:Grand Total 35
Equation 14: Grand Total mean 35
Equation 15: One-Way ANOVA model 36
Equation 16: Equation for Total Sum Square (SS_T) 37
Equation 17: Equation for Total variance due to treatment (SS_{TR}) 37
Equation 18: Two-Way ANOVA Model 40
Equation 19: Equation for Total Sum Square (SS_T) 41
Equation 20: Equation for Total variance due to treatment (SS_{TR}) 41
Equation 21:Finding Missing Value 44
Equation 22: Equation for Total Sum Square (SS_T) 47
Equation 23: Equation for Total variance due to treatment (SS_{TR}) 47

TABLES

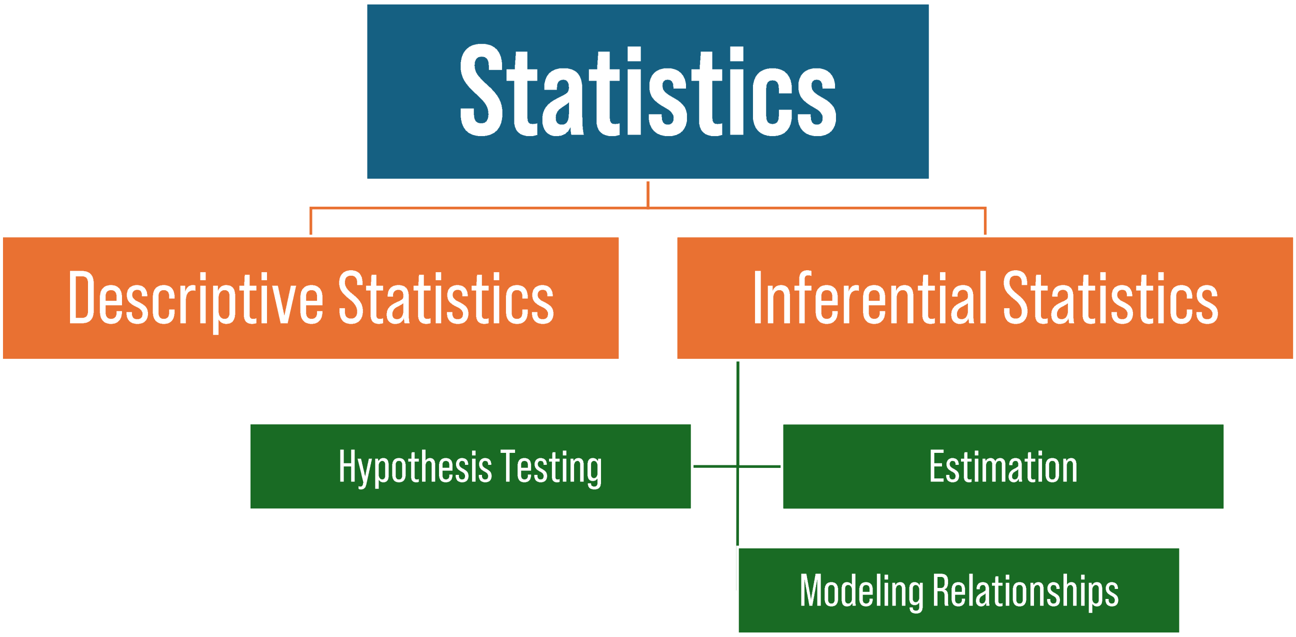
Table 1: One-Way ANOVA Table (Used for ANOVA related Calculation) 37
Table 2: Two-Way ANOVA Table (Used for ANOVA related Calculation) 41
Table 3:Latin Square Design: ANOVA Table 47

BASIC DEFINITIONS AND CONCEPTS

Kinds of Statistics

We can divide statistics in to two parts.

- Descriptive statistics
- Inferential statistics



KEY STATISTICAL CONCEPTS

Population

A population is the set of all the individuals of interest in a particular study.

It is an entire group of people or study elements, things or measurements having some common fundamental characteristics

Any actual or conceptual collection of individual items, defined by stranded characteristics.

Example:

- Advertisements for IT jobs in the Sri Lanka
- Songs from the VOICE Song Contest
- Undergraduate students in SUSL
- All countries of the world

Mainly the term population can be divided in to two parts.

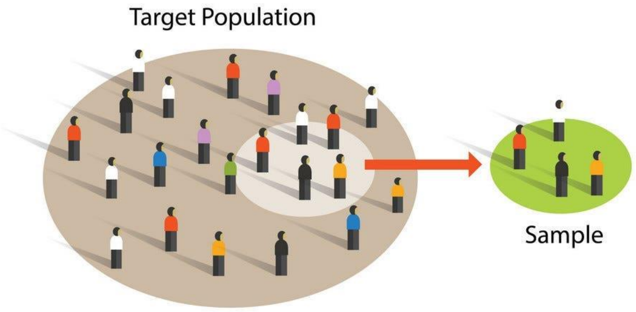
- Finite population
- Infinite population

Finite population: If a population consists of fixed number of values, then it is said to be finite. *Ex: Number of days per month.*

Infinite population: If a population consists of an endless succession of values, it is said to be infinite. *Ex: Number of insects in a certain region.*

Sample

A sample is a set of data drawn from the population (A sample is a small segment of the population). Potentially very large, but less than the population. In other words, a sample is a subset of a population.



Reasons for sampling

- **Necessity:** Sometimes it's simply not possible to study the whole population due to its size or inaccessibility.
- **Practicality:** It's easier and more efficient to collect data from a sample.
- **Cost-effectiveness:** There are fewer participant, laboratory, equipment, and researcher costs involved.
- **Manageability:** Storing and running statistical analyses on smaller datasets is easier and reliable

Collecting data from a sample

When your population is

- large in size,
- geographically dispersed, or difficult to contact,

it's necessary to use a sample. You can use sample data to make estimates or test hypotheses about population data.

Fundamentals methods for selecting sample

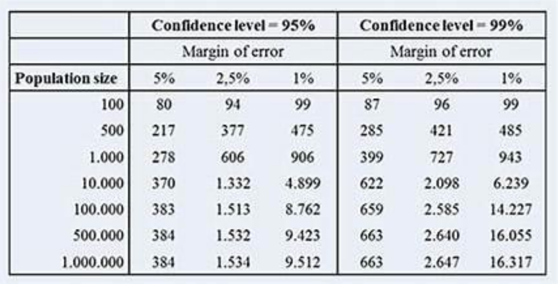
10% rule: Take at least 10% of the population as your sample.

The sample size at least needs to be 30 or more (So the population size needs to greater than 100 to use this rule)

If population size<100 we consider the population as sample.

Morgans table: Used to determine sample size.

If the population is very large go with the Morgans table **recommended**.



	Confidence level = 95%			Confidence level = 99%		
	Margin of error			Margin of error		
Population size	5%	2,5%	1%	5%	2,5%	1%
100	80	94	99	87	96	99
500	217	377	475	285	421	485
1.000	278	606	906	399	727	943
10.000	370	1.332	4.899	622	2.098	6.239
100.000	383	1.513	8.762	659	2.585	14.227
500.000	384	1.532	9.423	663	2.640	16.055
1.000.000	384	1.534	9.512	663	2.647	16.317

If the margin of error getting low the sample size will be increased

Formulas for determining the sample size:

$$Sample\,Size : \frac{\dfrac{z^2 \times p(1-p)}{e^2}}{1+\left(\dfrac{z^2 \times p(1-p)}{e^2 N}\right)}$$

$$N = \text{Population Size}$$
$$z = \text{Z-Score}$$
$$e = \text{Margine of error}$$
$$\text{p} = \text{Standard Deviation}$$

Equation 1: Formula for determining the sample size

Example:

- You want to study political attitudes in young people.
- Your population is the 30,000 undergraduate students in the Sri Lankan Universities.
- Because it's not practical to collect data from all of them, you use a sample of 300 undergraduate volunteers from three Sri Lankan universities
- This is the group who will complete your online survey.

Acceptance Sampling

Acceptance sampling is a statistical method used to decide whether to accept or reject a batch (or lot) of products. This inspection can occur at various points in the production process, including when products leave the factory or even during manufacturing.

There are two main types of inspection:

- 100% inspection (examining every item)
- Sampling inspection (examining only a sample)

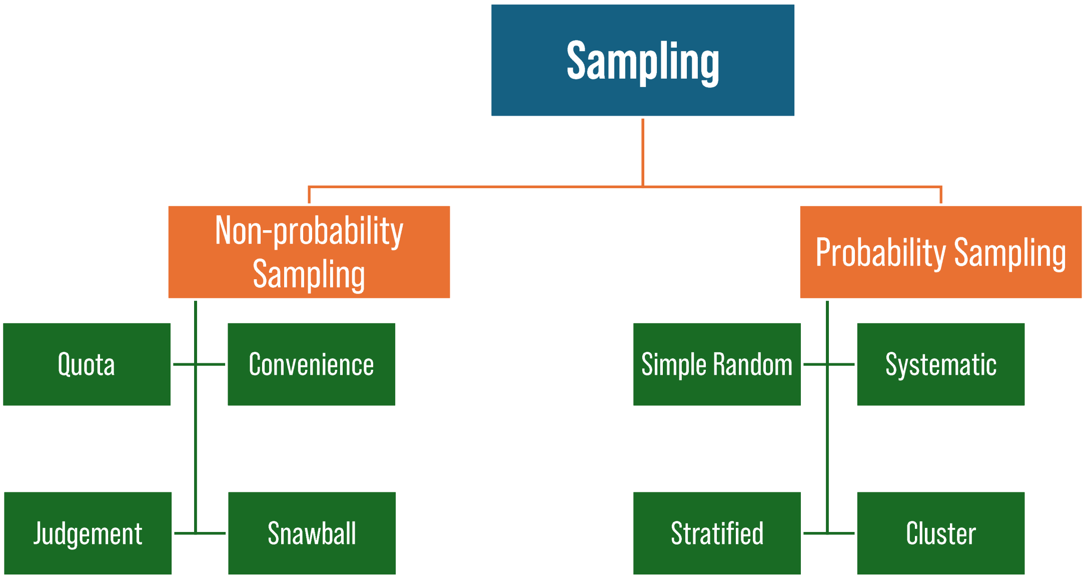
Most of the peoples are go with the Sampling Inspection because:

- The cost and time required for sampling inspection is quite less as compared to 100% inspection.
- Smaller inspection staff is necessary.
- Less damage to products because only few items are subjected to handling during inspection.
- The lot is disposed of in shorter time so that scheduling and delivery are important.

Sampling Techniques

Sampling Methods can be classified into one of two categories:

- Probability Sampling: Sample has a known probability of being selected
- Non-probability Sampling: Sample does not have known probability of being selected as in convenience or voluntary response surveys



Probability Sampling

In probability sampling it is possible to both determine which sampling units belong to which sample and the probability that each sample will be selected.

The following sampling methods are types of probability sampling:

- Simple Random Sampling (SRS)
- Stratified Sampling
- Cluster Sampling
- Systematic Sampling
- Multistage Sampling (in which some of the methods above are combined in stages)

Simple Random Sampling

Definition: A method of selecting n units from a population of size N such that every possible sample of size n has an equal chance of being selected.

Key Features:

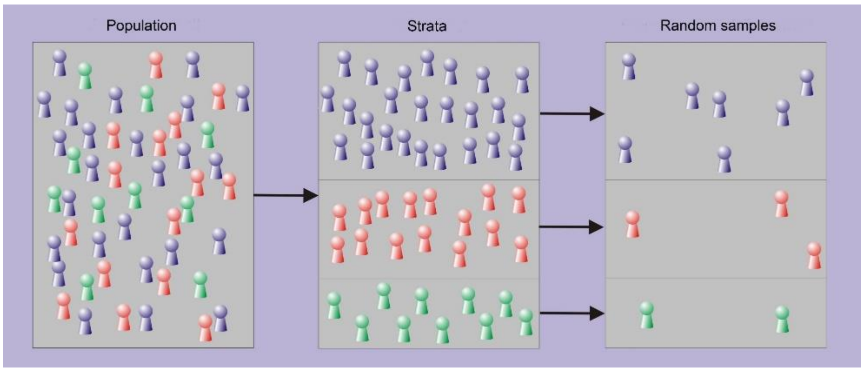
- Every element in the population has an equal opportunity of being included in the sample.
- Units in the population are numbered from 1 to N.
- A series of random numbers between 1 and N is generated using:
  - A table of random numbers, or
  - A computer program that produces random numbers.

Procedure:

- Number all units in the population from 1 to N.
- Randomly select the required number of units (n).
- If the population is divided into groups (strata), a simple random sample can be drawn independently from each group. This is called Stratified Random Sampling.

Stratified Sampling

Definition: A sampling method used when the population is not homogeneous but consists of homogeneous subgroups (called strata). The population is divided into non-overlapping subpopulations (N_1, N_2, ..., N_n) such that N1+N2+...N_N =N.



Key Features:

- Strata are created based on shared characteristics within each subgroup.
- A sample is drawn independently from each stratum.
- Sample sizes within strata are denoted as n_1, n_2, ..., n_n.
- If simple random sampling is used within each stratum, the procedure is called Stratified Random Sampling.

Example:

A school offers three streams: Science (135 students), Arts (45 students), Commerce (90 students). To select a sample of 30 students:

Calculate proportions:

Science: $\frac{135}{270}$ = 0.5,

Arts: $\frac{45}{270}$ = 0.167,

Commerce: $\frac{90}{270}$ = 0.333.

Allocate sample size proportionally:

Science: 30×0.5=15,

Arts: 30×0.167=5,

Commerce: 30×0.333=10.

Randomly select 15, 5, and 10 students from each stream using simple random sampling.

Systematic Sampling

Definition: A sampling method used for homogeneous populations when a complete list of items is unavailable. It involves selecting every k^th unit after randomly choosing a starting point.

Key Features:

Units in the population are numbered from 1 to N.

Interval (k) is calculated as:

$$\mathrm{k} = \frac{Population\ Size\ (N)}{Sample\ Size\ (n)}$$

If k is not an integer, round it to the nearest whole number.

Procedure:

- Determine interval k.
- Randomly select a starting point within the first k units.
- Select every kth unit thereafter until the sample size (n) is achieved.

Example:

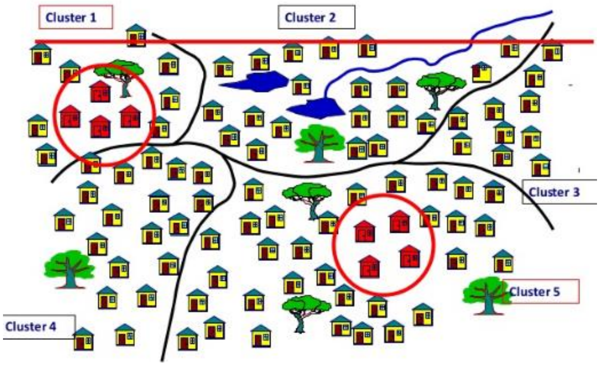
There are 150 students in a class, and we want a sample size of 20:

$$Calculate\ interval:\ k = \frac{150}{20} = 7.5 \approx 7$$

Randomly choose a starting point, ex: 2,
Select every 7th unit from 2: 2,9,16, until you have selected all 20 units.

Cluster Sampling

Definition: A sampling method where the population is divided into similar sub-groups (clusters), and entire clusters are randomly selected for study. Clusters should ideally be representative of the entire population.

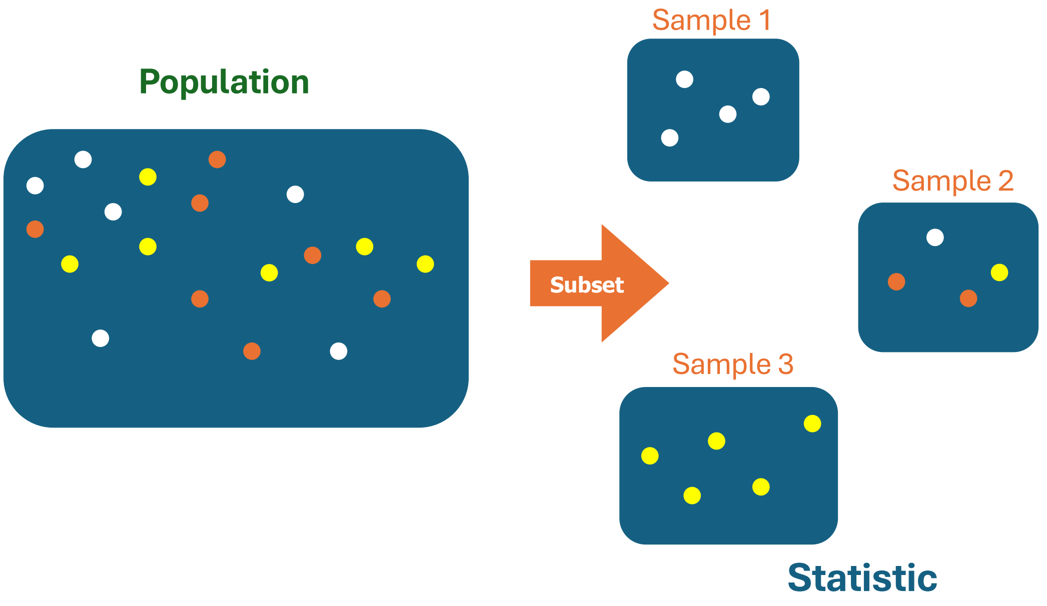


Key Features:

- Clusters are naturally occurring groups within the population (e.g., neighborhoods, schools).
- Instead of sampling individuals across all groups, entire clusters are sampled.

Example:

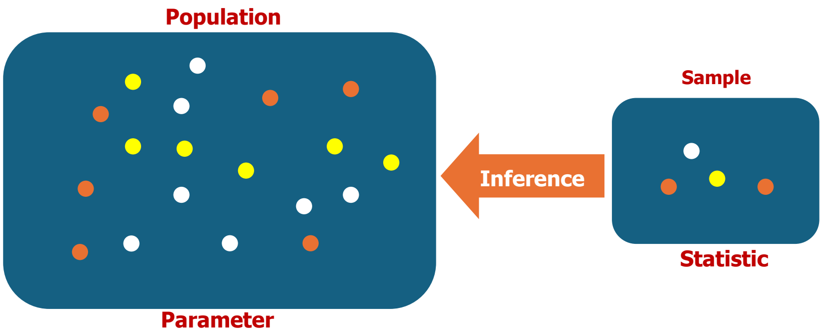
- To survey residents' opinions on public transport in a city divided into neighborhoods:
- Randomly select a few neighborhoods (clusters).
- Survey all residents within those neighborhoods.



Note: **Population has Parameters, Samples have Statistics.**

Statistical Inference

Statistical inference is the process of making an estimate, prediction, or decision about a population based on a sample.



Inferential statistics is used to draw conclusions or inferences about characteristics of populations based on data from a sample.

STATISTICAL ESTIMATION

Statisticians use sample statistics to estimate population parameters. For example, sample means are used to estimate population means; sample proportions, to estimate population proportions.

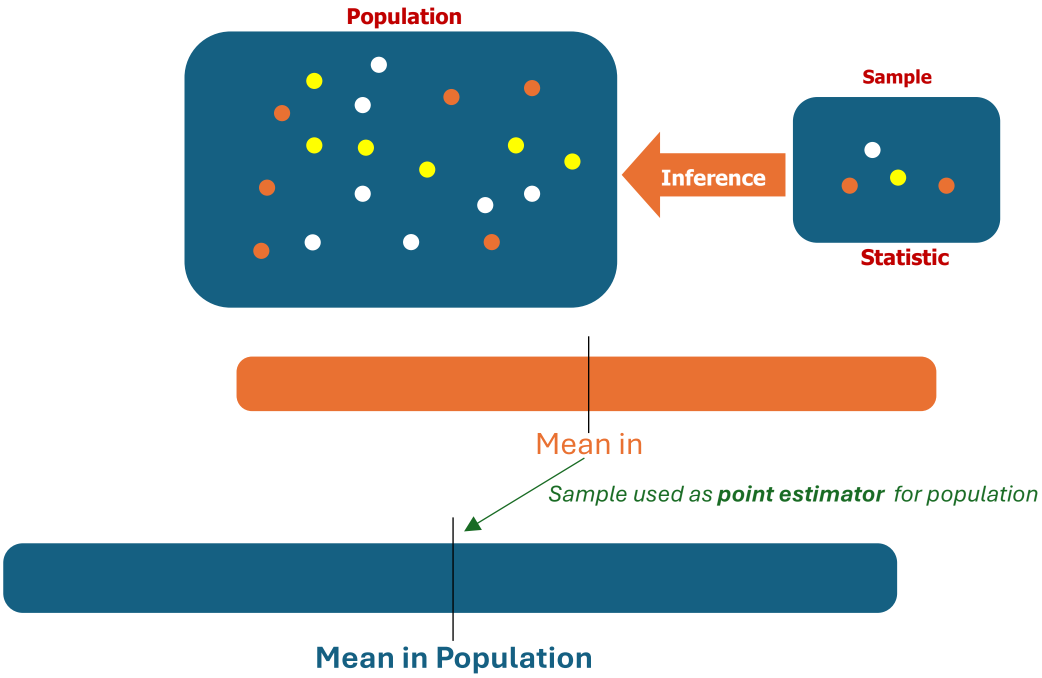
An estimate of a population parameter may be expressed in two ways:

- Point Estimation
- Interval Estimation

POINT ESTIMATION

A point estimate of a population parameter is a single value of a statistic.

For example, the sample mean $\bar{x}$ is a point estimate of the population mean and sample variance s^2 is a point estimation of the population variance σ^2.



Similarly, the sample proportion *p* is a point estimate of the population proportion *P*.

Properties of Point Estimators

Bias

The bias of a point estimator is defined as the difference between the expected value of the estimator and the value of the parameter being estimated.

When the **estimated value of the parameter** and **the value of the parameter being estimated** are equal, the estimator is considered unbiased.

Unbiased Estimates

A point estimate is a single value used to estimate a population parameter (e.g., mean or proportion). A point estimate is unbiased if its expected value equals the true value of the parameter:

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

Here:

$\hat{\theta}$: The point estimate.

θ: The true population parameter.

$E(\hat{\theta})$: The expected value of the point estimate.

Bias of an Estimate

If a point estimate is not unbiased, it has a bias. The bias is calculated as:

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

Bias measures how far the expected value of the point estimate is from the true parameter.

A smaller bias (closer to zero) means the point estimate is better.

Consistency

Consistency tells us how close the point estimator stays to the value of the parameter as it increases in size. The point estimator requires a large sample size for it to be more consistent and accurate. You can also check if a point estimator is consistent by looking at its corresponding expected value and variance.

Example:

Suppose that $E(X_1) = \mu,\ Var(X_1) = 10,\ E(X_2) = \mu,\ and\ Var(X_2) = 15,$ and consider the point estimates:

$$\mu_1 = \frac{X_1}{2} + \frac{X_2}{2},$$

$$\mu_2 = \frac{X_1}{4} + \frac{3X_2}{4},$$

$$\mu_3 = \frac{X_1}{6} + \frac{X_2}{3} + 9.$$

(a) Calculate the bias of each point estimate. Is any one of them unbiased?

(b) Calculate the variance of each point estimate. Which one has the smallest variance?

Answer:

$$\hat{\mu}_1 = \frac{X_1}{2} + \frac{X_2}{2}$$

$$E(\mu) = \frac{1}{2}E(X_1) + \frac{1}{2}E(X_2) = \frac{1}{2}\mu + \frac{1}{2}\mu = \mu$$

$$Bias = E(\mu_1) - \mu = 0$$

$$\hat{\mu}_2 = \frac{X_1}{4} + \frac{3X_2}{4}$$

$$E(\mu_2) = \frac{1}{4}E(X_1) + \frac{3}{4}E(X_2) = \frac{1}{4}\mu + \frac{3}{4}\mu = \mu$$

$$Bias = E(\mu_2) - \mu = 0$$

$$E(\mu_3) = \frac{1}{6}E(X_1) + \frac{1}{3}E(X_2) + 9 = \frac{1}{6}\mu + \frac{1}{3}\mu + 9 = \frac{\mu}{2} + 9$$

$Bias = E(\mu_3) - \mu = 9$ So, in here $\mu_1\ and\ \mu_2\ are\ unbiased$ but $\mu_3\ is\ Biased$

$$Var(\mu_1) = \frac{1}{4}Var(X_1) + \frac{1}{4}Var(X_2)$$
$$= 6.25$$

$$Var(\mu_2) = \frac{1}{16}Var(X_1) + \frac{9}{16}Var(X_2)$$
$$= 9.0625$$

$Var(\mu_1) < Var(\mu_2)$ *therefore* μ_1 *is smalles*

CONFIDENCE INTERVALS

What Are Confidence Intervals?

Confidence intervals are a range of values that estimate where the true population mean is likely to be. It shows how much error there might be between the sample mean (from data) and the actual population mean.

A confidence interval gives two limits:

- **Lower Limit:** The smallest value the population mean could be.
- **Upper Limit:** The largest value the population mean could be.

Why do we need Confidence Intervals?

Confidence intervals help us understand how close our sample data is to the true population value. Instead of guessing one number (point estimate), it gives a range where the true value is likely to lie.

Interval Estimation

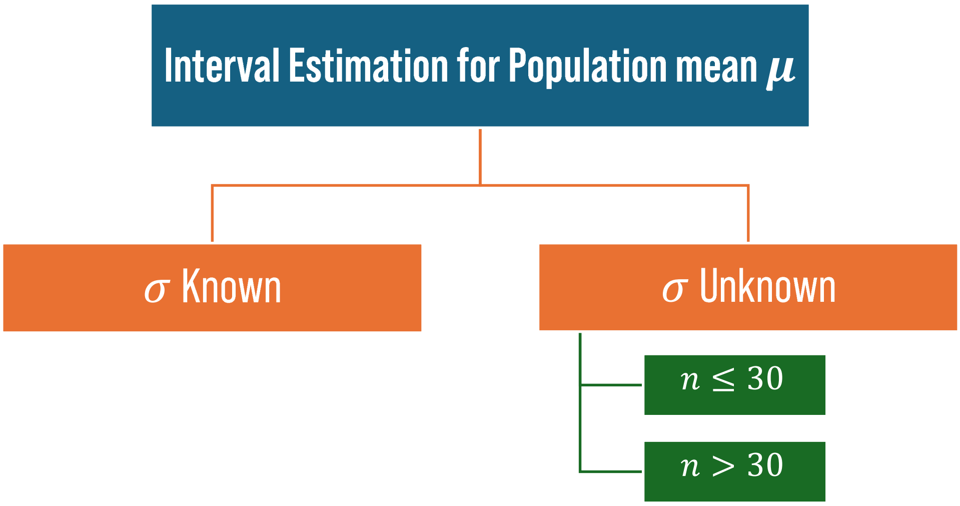
Interval estimation means finding two numbers (limits) between which the population parameter lies.

Example:
$a < \mu < b$
This means the population mean μ is greater than a and less than b.

How to Estimate Confidence Intervals for Sample Mean

Conditions Needed:

- The sampling method must be simple random sampling.
- The sampling distribution (data pattern) must be approximately normal. This happens when:
  - The population itself is normally distributed.
  - The sample size is ≤ 30, and the data is symmetric, unimodal, and has no outliers.
  - The sample size is > 30, and there are no outliers.



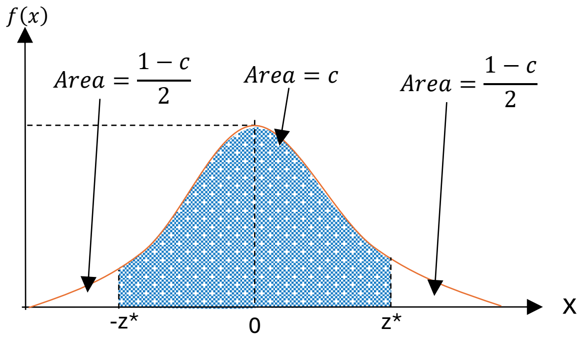
Interval Estimation for population mean μ when σ^2 is known

Let $x_1, x_2, x_3, \ldots, x_n$ be a random sample of size n from a normal distribution with mean μ and variance σ^2.

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

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Equation 2: Equation for determine confidence Interval



Example:

According to a random sample of 30 students of a certain institute, the average reading speed of the students is 80 words per minute with a population standard deviation 6.8 words per minute.

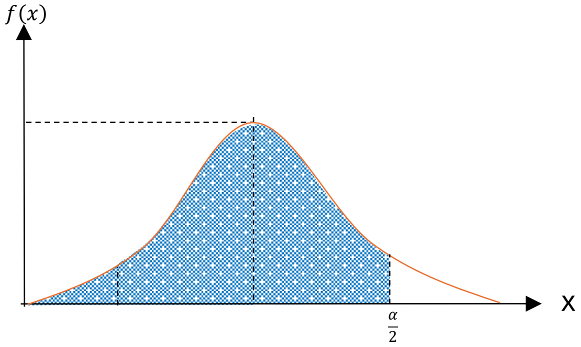
Constrict 68%, 95% and 99% confidence interval for the mean reading speed of a student of the institute.

Answer:

$(1-\alpha)\%$ Confidence Interval for μ when σ known...

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$\bar{x} = 80,\ \sigma = 6.8,\ n = 30$



$(1-\alpha)\%$ Confidence Interval for μ

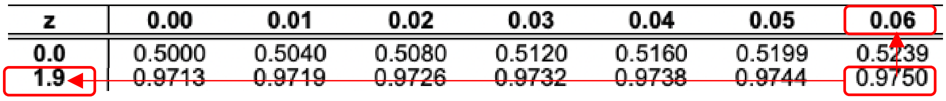
$1-\alpha = 0.95$

$\alpha = 0.05$

$\frac{\alpha}{2} = 0.025$

So for the rest of the area $= 1 - \frac{\alpha}{2}$

$= 0.975$



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750

So for $Z_{0.975} = 1.96$

So,

$$\text{Confidence Interval for } \mu = \bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$$= 80 \pm 1.96 \times \frac{6.8}{\sqrt{30}}$$

$$= 80 \pm 2.4334$$

$$= 77.5666 \le \mu \le 82.4334$$

Example:

The fertilizer mixing machine is set to add Nitrite for every bag of fertilizer. Randomly selected 16 bags are examined. The percentages of Nitrate of the bags are as below.

21.8	21.6	21.0	20.9	19.8	19.6	20.9	21.1
20.4	20.6	19.7	19.6	20.3	23.7	20.5	20.8

It is given that the population standard deviation of the distribution is 0.48. Find the 95% confidence interval for the true mean percentage of Nitrate of the distribution.

We are 95% confident that the population of all bags.

Answer:

$(1-\alpha)\%$ Confidence Interval for μ when σ known...

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$\sigma = 0.48,\ n = 16$

$$\bar{x} = \frac{332.3}{16} = 20.76875$$

$(1-\alpha)\%$ Confidence Interval for μ

$1-\alpha = 0.95$

$\alpha = 0.05$

$\frac{\alpha}{2} = 0.025$

So for the rest of the area = $1-\frac{\alpha}{2}$

$= 0.975$

So for $Z_{0.975} = 1.96$

So,

$$\text{Confidence Interval for } \mu = \bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$$= 20.76875 \pm 1.96 \times \frac{0.48}{\sqrt{16}}$$

$$= 20.76875 \pm 0.2352$$

$$= 20.53355 \le \mu \le 21.00395$$

Interval Estimation for Population mean μ when σ^2 is unknown and n $\leq$ 30

Frequently, we are attempting to estimate the mean of a population when the variance is unknown. If we have a random sample from a normal distribution, then the random variable T,

$T = \frac{\bar{X}-\mu}{S/\sqrt{n}}$ has a student t-distribution with n-1 degrees of freedom.

Let x_1, x_2, x_3, ..., x_n be a random sample of size n from a normal distribution with mean μ and variance σ^2

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

then,

$$P\left(-t_{\alpha/2} < T < t_{\alpha/2}\right) = 1-\alpha$$

$$P\left(-t_{\alpha/2} < \frac{\bar{X}-\mu}{S/\sqrt{n}} < t_{\alpha/2}\right) = 1-\alpha$$

Then $100(1-\alpha)\%$ confidence interval for μ is,

$$\left(\bar{x} - t_{\alpha/2}\frac{S}{\sqrt{n}} < \mu < \bar{x} - t_{\alpha/2}\frac{S}{\sqrt{n}}\right)$$

Where $t_{\alpha/2}$ is the t-value with v=n-1 degree of freedom

$$\bar{x} \pm t_{(\alpha/2, n-1)}\frac{S}{\sqrt{n}}$$

Equation 3: Equation for determine confidence Interval when we don't know σ

Example:

Let *X* be the weight in grams of a 52-gram snack pack of candies. Assume that *X*, is normally distributed. The weights of 10 snack packs are;

55.54	56.54	57.58	55.13	57.48
56.06	59.93	58.30	52.57	58.46

Find a 95% confidence interval for the mean percentage of candies.

Answer:

$(1-\alpha)\%$ Confidence Interval for μ when σ unknown...

$$\bar{x} \pm t_{(\alpha/2, n-1)}\frac{S}{\sqrt{n}}$$

X	$(\bar{X}-X)$	$(\bar{X}-X)^2$
52.57	4.189	17.547721
55.13	1.629	2.653641
55.54	1.219	1.485961
56.06	0.699	0.488601
56.54	0.219	0.047961
57.48	-0.721	0.519841
57.58	-0.821	0.674041
58.3	-1.541	2.374681
58.46	-1.701	2.893401
59.93	-3.171	10.055241
AVERAGE=56.759		SUM=38.74109

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

$$S=\sqrt{\frac{38.74109}{9}}$$

$$S=2.075,\ n=10$$

$$\bar{x}=\frac{567.6}{10}=56.76$$

$(1-\alpha)\%$ Confidence Interval for μ

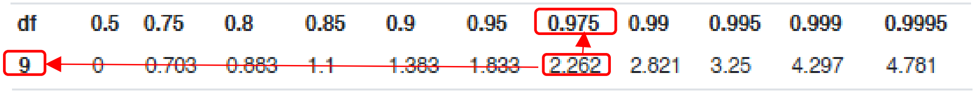
$$1-\alpha=0.95$$

$$\alpha=0.05$$

$$\frac{\alpha}{2}=0.025$$

So for the rest of the area = $1-\frac{\alpha}{2}$

$$=0.975$$



df	0.5	0.75	0.8	0.85	0.9	0.95	0.975	0.99	0.995	0.999	0.9995
9	0	0.703	0.883	1.1	1.383	1.833	2.262	2.821	3.25	4.297	4.781

So here $t_{(\alpha/2,n-1)}\rightarrow t_{(0.975,9)}$

So the Critical Value for here$=2.262$

So,

$$\text{Confidence Interval for }\mu=\bar{x}\pm t_{(\alpha/2,n-1)}\frac{S}{\sqrt{n}}$$

$$=56.76\pm2.262\times\frac{2.075}{\sqrt{10}}$$

$$=56.76\pm1.484$$

$$=55.276\leq\mu\leq58.244$$

Interval Estimation for Population mean μ when σ^2 is unknown and n $>$ 30

In some cases n is too much large we can't find that much of n value in t-distribution Table at that time,

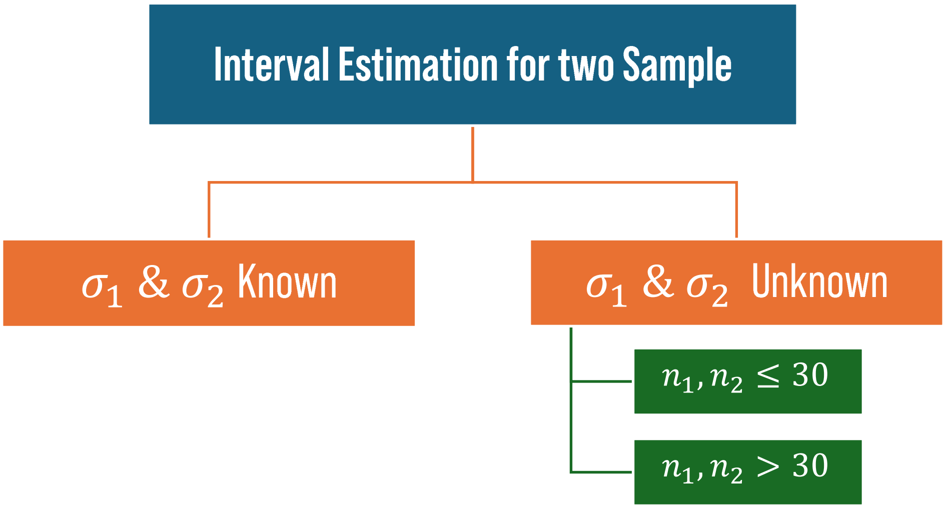
$$t_{(\alpha/2, n-1)} \to z_{\alpha/2}$$

$$\bar{x} \pm z_{\alpha/2} \frac{S}{\sqrt{n}}$$

Equation 4: Equation for determine confidence Interval when we don't know σ and n value too large as well

Interval Estimation for two samples

In here we estimate the Difference between Two Means...



Interval Estimation for two Sample

σ_1 & σ_2 Known

σ_1 & σ_2 Unknown

$n_1, n_2 \leq 30$

$n_1, n_2 > 30$

Estimation Requirements

This approach is valid whenever the following conditions are met:

- Both samples are simple random samples.
- The samples are independent.
- Each population is at least 10 times larger than its respective sample.
- The sampling distribution of the difference between means is approximately normally distributed.

Two Samples: Estimating the Difference between Two Means

If we have two populations with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2. For the two independent random samples, one from each population of size n_1 and n_2, then the sampling distribution of $\bar{X}_1 - \bar{X}_2$ then,

$$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}\right)$$

$$\left(\bar{x} - \bar{y}\right) \pm z_{(\alpha/2)} \sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}$$

Equation 5: Estimating the Difference between Two Means

Example:

Two companies A and B produce same kind of nails. It is interested to find the confidence interval for the difference between the mean lengths of nails.

Let X be the length of a nail company A and Y be the length of a nail of company B. Suppose that we are taking a sample of 30 nails from A and a sample of 25 nails from B. It shows that $\bar{X} = 300mm$ and $\bar{Y} = 250mm$.

According to the production reports of the last year the variance of the lengths of nails of company A is $\sigma_A^2 = 400mm^2$ and that of company B is $\sigma_B^2 = 425mm^2$.

Assuming that the data are normally distributed, find 95% ,99% confidence interval for $\mu_A - \mu_B$, where μ_A and μ_B are population means of A and B respectively.

$(1-\alpha)\%$ Confidence Interval for μ when σ known...

$(1-\alpha)\%$ Confidence Interval for μ

$1-\alpha = 0.95$

$\alpha = 0.05$

$\frac{\alpha}{2} = 0.025$

So for the rest of the area = $1-\frac{\alpha}{2}$

$= 0.975$

So for $Z_{0.975} = 1.96$

So,

$$\left(\overline{X} - \overline{Y}\right) \pm z_{(\alpha/2)} \sqrt{\frac{\sigma_1{}^2}{n_1} + \frac{\sigma_2{}^2}{n_2}}$$

$\sigma_A{}^2 = 400mm^2, \sigma_B{}^2 = 425mm^2, n_A = 30, n_B = 25, \overline{X} = 300mm, \overline{Y} = 250mm$

$$\text{Confidence Interval for } \mu_A - \mu_B = \left(\overline{X} - \overline{Y}\right) \pm z_{(\alpha/2)} \sqrt{\frac{\sigma_A{}^2}{n_A} + \frac{\sigma_B{}^2}{n_B}}$$

$$= \left(300 - 250\right) \pm 1.96 \times \sqrt{\frac{400}{30} + \frac{425}{25}}$$

$$\text{Confidence Interval for } \mu_A - \mu_B = 39.22 \le \mu_A - \mu_B \le 60.78$$

Interval Estimation of difference of the means of Two Populations when σ_1, σ_2 are Unknown and $n_1, n_2 \le 30$

In here we don’t know the Population variance of both samples so, we find a “pooled estimate of the population standard deviation” for both samples.

$$S_p{}^2 = \frac{S_1{}^2(n_1 - 1) + S_2{}^2(n_2 - 1)}{(n_1 + n_2 - 2)}$$

S_1^2=Sample Variance for Sample 1

S_2^2= Sample Variance for Sample 2

Equation 6: pooled estimate of the population standard deviation

$$\left(\bar{x}-\bar{y}\right) \pm t_{(\alpha/2, n_1+n_2-2)} S_p \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

Equation 7:Equation for determine confidence Interval Difference when we don't know σ

Example:

It is claimed that a new diet will reduce a person's weight by 4.5 kilograms on the average in a period of 2 weeks. The weight of 7 women two groups who followed this diet were recorded.

Test a manufacturer's claim by computing a 95% confidence interval for the mean difference in their weight.

Assume that the count differences of weights to be approximately normally distributed

Women	Group 01	$(X_i - \bar{X}_{G1})^2$	Group 02	$(X_i - \bar{X}_{G2})^2$
1	58.5	11.08	60.0	2.99
2	60.3	2.34	54.9	11.37
3	61.7	0.02	58.1	0.03
4	69.0	51.43	62.1	14.66
5	64.0	4.72	58.5	0.05
6	62.6	0.60	59.9	2.65
7	56.7	26.30	54.4	14.99
	AVG = 61.83	**SUM =96.47**	**AVG = 58.27**	**SUM = 46.73**

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

$$S_1 = 4.01$$

$$S_2 = 2.79$$

$$S_p{}^2 = \frac{S_1{}^2(n_1-1)+S_2{}^2(n_2-1)}{(n_1+n_2-2)}$$

$$S_p{}^2 = \frac{4.01^2\times(7-1)+2.79^2\times(7-1)}{(14-2)}$$

$$S_p{}^2 = 11.9321$$

$$S_p = 3.4543$$

So,

$$n_1 = 7, n_2 = 7,\ \bar{X} = 61.83kg,\ \bar{Y} = 58.27kg$$

$$\text{Confidence Interval for } \mu_A - \mu_B = \left(\bar{x}-\bar{y}\right) \pm t_{(\alpha/2,n_1+n_2-2)} S_p \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

$$= 61.83 - 58.27 \pm 2.179 \times 3.4543 \times \sqrt{\frac{1}{7}+\frac{1}{7}}$$

$$\text{Confidence Interval for } \mu_A - \mu_B = -0.4633 \le \mu_A - \mu_B \le 7.5833$$

$(1-\alpha)$% Confidence Interval for μ when σ known...

$(1-\alpha)$% Confidence Interval for μ

$$1-\alpha = 0.95$$

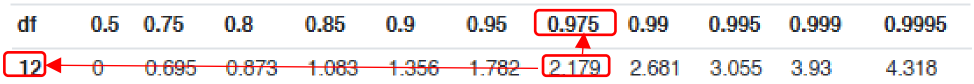
$$\alpha = 0.05$$

$$\frac{\alpha}{2} = 0.025$$

So for the rest of the area = $1 - \frac{\alpha}{2} = 0.975$

So here $t_{(\alpha/2,n_1+n_2-2)} \rightarrow t_{(0.975,12)}$

So the Critical Value for here $= 2.179$



df	0.5	0.75	0.8	0.85	0.9	0.95	0.975	0.99	0.995	0.999	0.9995
12	0	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.93	4.318

Figure 1: t-distribution Table for one tail

CONFIDENCE INTERVAL FOR A POPULATION PROPORTION

In here we find population proportion from Sample proportion...

If X is a binomial random variable;

$$X \sim B(n,\ p)$$

where n is the number of trials and p is the probability of a success.

To form a sample proportion,

$$P' = \frac{X}{n}$$

- P' = the estimated proportion of successes or sample proportion of successes
- X = the number of successes in the sample
- n = the size of the sample

This lesson describes how to construct a confidence interval for a sample proportion, p, when the sample size is large.

Estimation Requirements

- The sampling method is simple random sampling.
- The sample should sufficiently large. As a **rule of thumb**.
  - $np \geq 10$
  - $nq \geq 10$

Single Sample: Estimating a Proportion

If $\hat{p}$ is the proportion of successes in a random sample of size n, and $\hat{q} = 1 - \hat{p}$,

$$then, Z = \frac{P - p}{\sqrt{\frac{pq}{n}}}$$

$$Assumption: np \geq 10 \ \& \ nq \geq 10$$

$$Hence,$$

$$P\left(-z_{\alpha/2} < z < z_{\alpha/2}\right)$$

$$P\left(-z_{\alpha/2} < \frac{P - p}{\sqrt{\frac{pq}{n}}} < z_{\alpha/2}\right)$$

An approximate 100(1-α)% CI for the binomial parameter p is given by...

$$\boxed{P \pm z_{\alpha/2}\sqrt{\frac{pq}{n}}}$$

Equation 8: Confidence Interval for Population Proportion

Example:

In the year 2016 Youth Risk Behavior survey done by the Research Centers for Disease Control, 747 out of *n* = 1168 female graders said the always use a seatbelt when driving.

Goal: Estimate proportion always using seatbelt when driving in the population of grade female drivers.

Checking Requirenment

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

$$= \frac{747}{1168}$$

$$p = 0.6396$$

$$np = 1168 \times 0.6396 = 747$$

$$np \geq 10$$

$$similarly$$

$$nq = n(1-p) = 1168 \times 0.3104 = 362.55$$

$$nq \geq 10$$

Requirenment fulfilled

$$1-\alpha = 0.95$$

$$\alpha = 0.05$$

$$\frac{\alpha}{2} = 0.025$$

So for the rest of the area = $1\text{-}\frac{\alpha}{2}$

$$= 0.975$$

So for $\text{Z}_{0.975}$ =1.96

So,

$$CI = \widehat{p} \pm z_{\alpha/2}\sqrt{\frac{pq}{n}}$$

$$\text{C.I for Proportion } (p) = 0.6396 \pm 1.96 \times \sqrt{\frac{0.6396 \times 0.3104}{1168}}$$

$$= 0.616 \leq p \leq 0.665$$

Compare the two proportions by finding Difference in Popular Proportion

$$\boxed{(p_1 - p_2) \pm z_{\alpha/2}\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}}$$

Equation 9: C.I. for the difference in two population proportions

Example:

Would you date someone with a great personality even though you did ***not*** find them attractive?

- **Women**: 0.611 of 131 answered “yes.”
- **Men**: 0.426 of 61 answered “yes.”

Confidence interval for the difference in *population proportions* of women and men who would say yes.

For 95% C.I...

$$1-\alpha = 0.95$$
$$\alpha = 0.05$$
$$\frac{\alpha}{2} = 0.025$$

$$\text{So for the rest of the area} = 1\text{-}\frac{\alpha}{2}$$
$$= 0.975$$

So for $Z_{0.975} = 1.96$

So,

$$\text{CI Difference for Proportion} = (p_1 - p_2) \pm z_{\alpha/2}\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

$$\text{CI Difference for Proportion} = (0.611 - 0.426) \pm 1.96 \times \sqrt{\frac{0.611(1-0.611)}{131} + \frac{0.426(1-0.426)}{61}}$$
$$= 0.0355 \le p \le 0.3345$$

So, A 95% confidence interval for the difference is 0.0355 to 0.3345 or 3.55% to 33.45%.

Single Sample: Estimating the Variance

If a sample size of n is drawn from a normal population with variance σ^2, and the sample variance s^2 is computed, then,

$X^2 = \frac{(n-1)s^2}{\sigma^2}$ has a chi-squared distribution with n-1 degrees of freedom when samples are chosen from a normal population.

Then, $(1-\alpha)\%$ C.I for σ^2 is...

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}$$

$$\sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2}}} \le \sigma \le \sqrt{\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}}$$

Equation 10: Confidence Interval for Variance & Standard Deviation of population

Example:

A statistician chooses 27 randomly selected dates, and when examining the occupancy records of a particular motel for those dates, finds a standard deviation of 5.86 rooms rented.

If the number of rooms rented is normally distributed, find the 95% confidence interval for the population standard deviation of the number of rooms rented.

For a sample size of n=27, we will have $df = n - 1 = 26$ degrees of freedom.

For a 95% confidence interval, we have $\propto = 0.05$, which gives 2.5% of the area at each end of the chi-square distribution.

$(1-\alpha)\%$Confidence Intervel for α^2 is...

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}$$

$1-\alpha = 0.95$

$\alpha = 0.05$

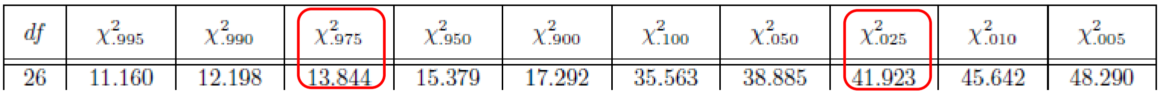
$\frac{\alpha}{2} = 0.025$

$1-\frac{\alpha}{2} = 0.975$

According to the CHI-Square Table for df=26 $(df = n-1)$

$\chi^2_{0.025} = 41.923$

$\chi^2_{0.975} = 13.844$



df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290

Figure 2: Chi-Square Distribution Table

So,

The Confidence Interval for Variance,...

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}$$

$$\frac{(27-1)\times 5.86^2}{41.923} \le \sigma^2 \le \frac{(27-1)\times 5.86^2}{13.844}$$

$21.297 \le \sigma^2 \le 64.492$

Example:

The following are the weights, in decagrams, of 10 packages of grass seed distributed by a certain company: 46.4, 46.1, 45.8, 47.0, 46.1, 45.9, 45.8, 46.9, 45.2, and 46.0. find a 95% CI for the variance of all such packages of grass seed distributed by this company, assuming a normal population.

Package	Weights(decagram)	$(X_i - \bar{X})^2$
1	46.4	0.078
2	46.1	0.000
3	45.8	0.102
4	47.0	0.774
5	46.1	0.000
6	45.9	0.048
7	45.8	0.102
8	46.9	0.608
9	45.2	0.846
10	46.0	0.014
	AVG = 46.12	**SUM =2.576**

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

$$S^2 = \frac{2.576}{9} = 0.286$$

$(1-\alpha)\%$Confidence Intervel for α^2 is...

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}$$

$$1-\alpha = 0.95$$

$$\alpha = 0.05$$

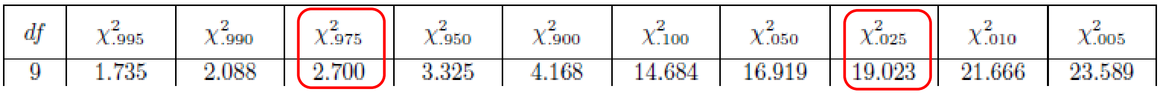
$$\frac{\alpha}{2} = 0.025$$

$$1-\frac{\alpha}{2} = 0.975$$

According to the CHI-Square Table for df=9 $(df = n-1)$

$$\chi^2_{0.025} = 19.023$$

$$\chi^2_{0.975} = 2.7$$



df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589

So,

The Confidence Interval for Variance,...

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}$$

$$\frac{(10-1)\times 0.286}{19.023} \le \sigma^2 \le \frac{(10-1)\times 0.286}{2.7}$$

$$0.1353 \le \sigma^2 \le 0.9533$$

$$0.3678 \le \sigma \le 0.9764$$

HYPOTHESIS TESTING

What is Statistical Hypothesis Testing?

It is a **method to make decisions** using data. Helps determine which of two or more claims (hypotheses) about a population is **more likely** to be true.

Why is it Important?

- Used to **test assumptions** about a population using sample data.
- Makes decisions **more accurate** than just looking at data.

Two Main Types of Statistical Inference:

- Estimation of population parameters
- Hypothesis Testing (focus here)

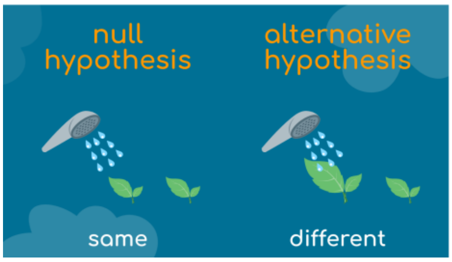
Examples of Statistical Hypothesis Testing:

- A new medicine is better than the old one.
- Men are taller than women.
- One toothpaste whitens teeth better than another.
- An ad works better on the right side of a page than the left.

What is a Statistical Hypothesis?

It is an **assumption about a population parameter** (e.g., mean, proportion). If sample data are not consistent with the statistical hypothesis, the hypothesis is rejected. There are two types of statistical hypotheses.

- Null Hypothesis
- Alternate Hypothesis



Null Hypothesis (H_0):

This is the **default** or starting belief — like saying,

"Nothing special is going on."

It assumes **there is no difference** between what you're comparing.

Ex:

Figure 3: Cake I

Figure 4: Cake II

H_0: There is **no difference in taste** between Cake I and Cake II.

You're saying both cakes taste the same — any difference you taste is just by **chance**.

Alternative Hypothesis (H_1 or H_a):

This is the opposite of H_0. You believe:

"Something is going on."

It assumes **there is a real difference** between what you're comparing.

Ex:

Figure 5: Cake I

Figure 6: Cake II

H_1: There **is a difference in taste** between Cake I and Cake II.

You're saying one cake **does** taste different from the other — and it's **not just chance**.

"Alternative Hypothesis is EXPECTED because if we usually change anything we expect the result should be also change"

Ex:

Let's say a company makes a **new fertilizer**. They believe it'll help plants grow faster.

- **H_0:** The new fertilizer has no effect — plants grow just like before.
- **H_1:** The new fertilizer does affect plant growth — maybe faster.

Example:

You which to show that μ exceeds 50.
$H_0: \mu \leq 50 \; VS \; H_1: \mu > 50$

You wish to show that μ is less than 50.
$H_0: \mu \geq 50 \; VS \; H_1: \mu < 50$

You wish to find evidence against the claim that μ is 50 when you believe it is greater than 50.
$H_0: \mu \leq 50 \; VS \; H_1: \mu > 50$

You wish to find evidence against the claim that μ is 50 when you believe it is less than 50.
$H_0: \mu \geq 50 \; VS \; H_1: \mu < 50$

HYPOTHESIS TESTS

Statisticians follow 4 main steps to test an idea (a hypothesis) using data:

Step 1: State the Hypotheses

- Write down two statements:
  - **Null Hypothesis (H_0):** No effect or no difference.
  - **Alternative Hypothesis (H_1):** There is an effect or a difference.
- They can't both be true — only one can be correct.

Step 2: Make a Plan

- Decide **which method/test** you'll use to analyze the data.
- Choose your **test statistic** (like z-score, t-score, etc.).

Step 3: Use the Sample Data

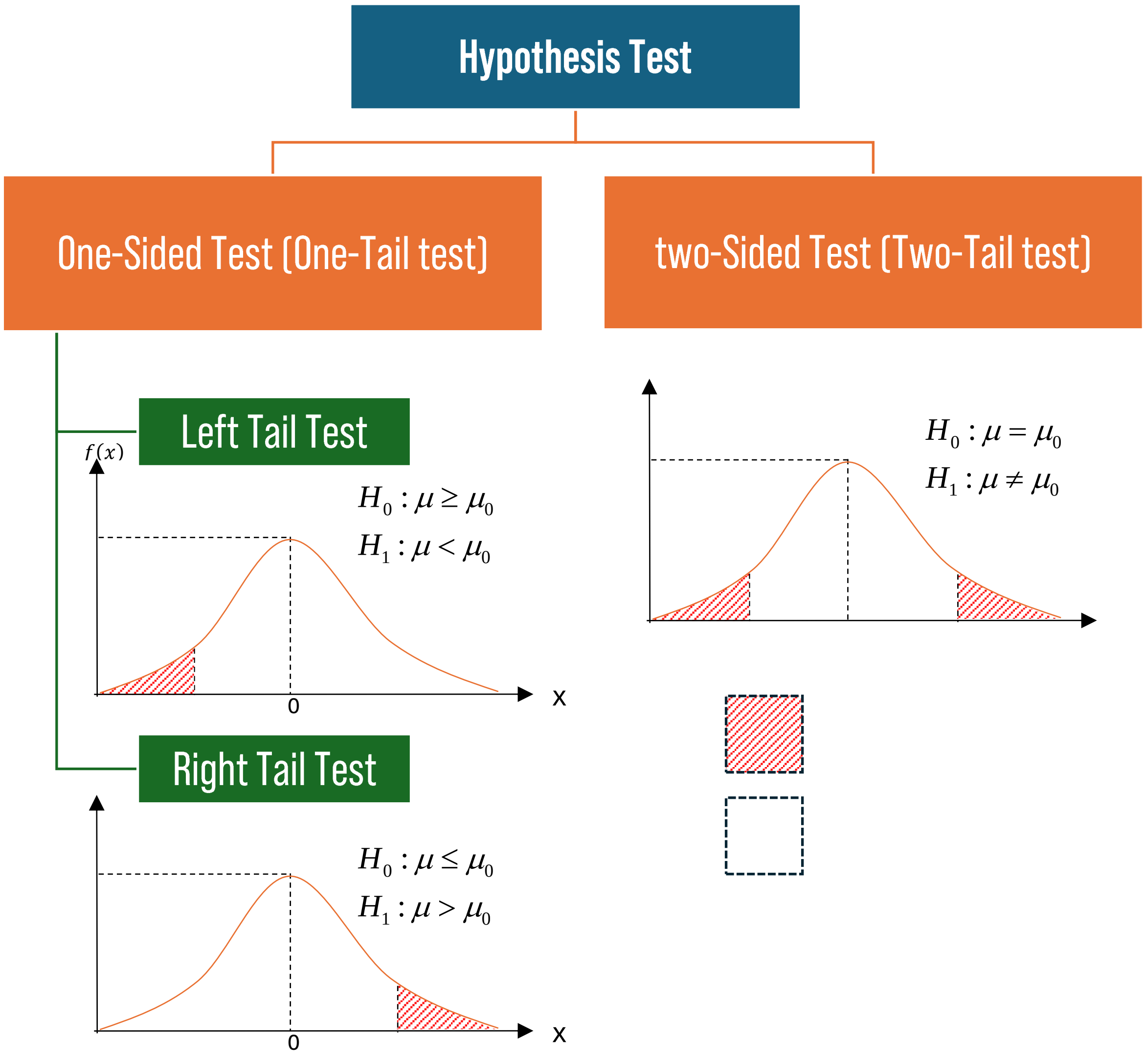
- Use the data you collected to calculate the test statistic.

Step 4: Make a Decision

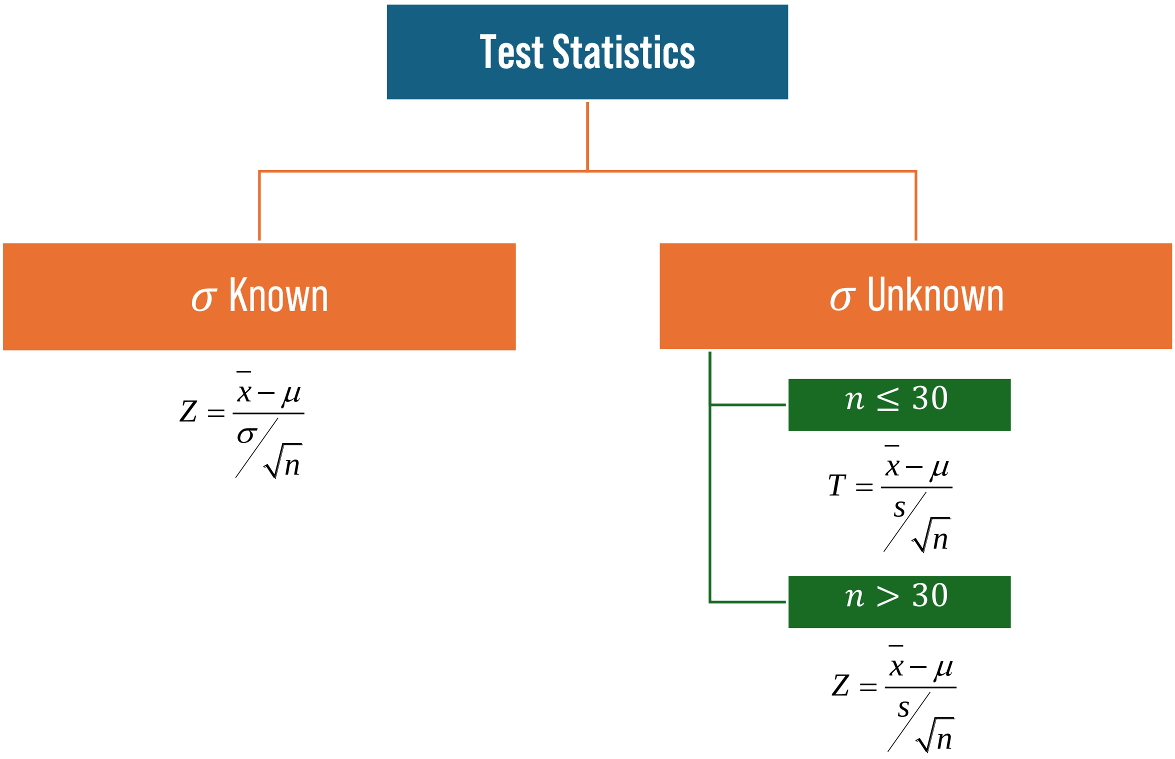
- Check if the result is likely or unlikely under the null hypothesis.
  - If **unlikely,** ❌ **reject H_0** → accept that something is happening.
  - If **likely,** ✅ **keep H_0** → no strong evidence of a change.

There are five ingredients to any statistical test

1. **Null Hypothesis**
2. **Alternate Hypothesis**
3. **Test Statistic**
4. **Rejection/Critical Region**
5. **Conclusion**



Test Statistics



Test Statistics

σ Known

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

σ Unknown

$n \leq 30$

$$T = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

$n > 30$

$$Z = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

Rejection/Critical Region

Significance level (α) =0.05

One-Sided Critical region

$$\alpha = 0.05$$

$$Z_{\alpha} = Z_{0.05} \Leftrightarrow Z_{0.95} = 1.65$$

$$OR$$

$$T_{\alpha,(n-1)} = T_{0.05,(n-1)} \Leftrightarrow T_{0.95,(n-1)}$$

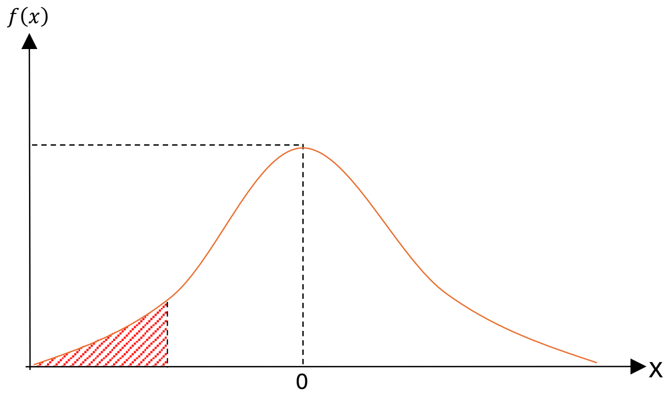


Figure 7: Left Tail Test

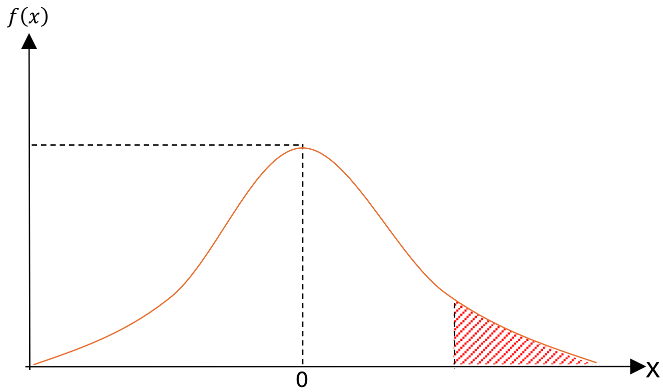
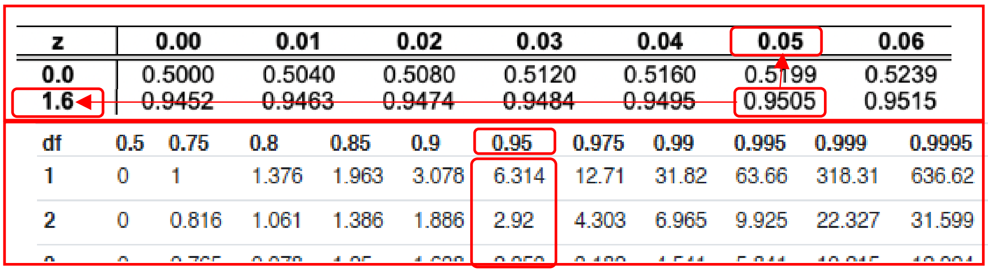


Figure 8: Right tail Test



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515

$\leftarrow Z - Ditribution$

df	0.5	0.75	0.8	0.85	0.9	0.95	0.975	0.99	0.995	0.999	0.9995
1	0	1	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0	0.816	1.061	1.386	1.886	2.92	4.303	6.965	9.925	22.327	31.599

$\leftarrow T - Ditribution$

Two-sided critical region

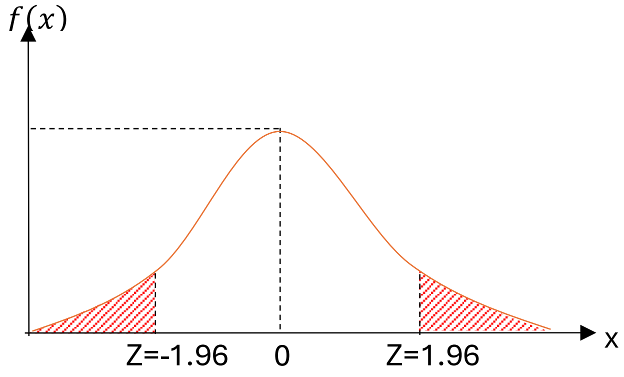
$$\alpha = 0.05$$

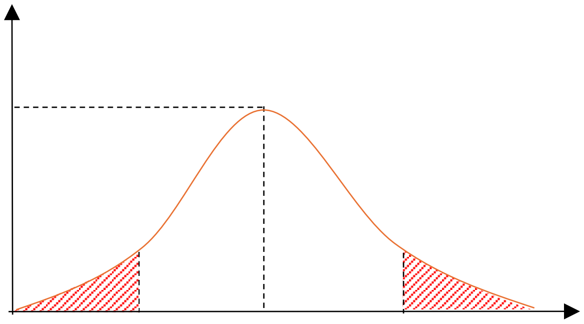
$$\frac{\alpha}{2} = 0.025$$

$$Z_{\frac{\alpha}{2}} = Z_{0.025} \Leftrightarrow Z_{0.975} = 1.96$$

$$OR$$

$$T_{\frac{\alpha}{2},(n-1)} = T_{0.025,(n-1)} \Leftrightarrow T_{0.975,(n-1)}$$





Conclusions

▨

□ H_0 Accepted Area (H_1 Rejected area)

Example: σ Known & Two Tailed test

Suppose that it is known from experience that, the population variance of the weight of 8-ounce packages of cookies made by a certain bakery is 0.16^2. **To check whether its production is under control on a given day, that is to check whether the true average weight of packages is 8 ounces.**

Employees select a random sample of 25 packages and find that their mean weight is $\bar{x} = 8.091$ ounce. Test the **alternative hypothesis $\boldsymbol{\mu \neq 8}$ against the, null hypothesis $\boldsymbol{\mu = 8}$** at the 0.05 level of significance.

Answer:

Step 01: Defining Hypothesis Testing.

$H_0: \mu = 8 \quad H_1: \mu \neq 8$

Step 02: Test Statistics

In here,

- we know population Variance $\sigma^2 = 0.16^2$

So, the Test Statistics: **Z**

Formula: $Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$

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

$$Z = \frac{8.091 - 8}{0.16 / \sqrt{25}} = 2.844$$

<u>Step 03:</u> Conclusion

$H_0 : \mu = \mu_0$

$\boxed{H_1 : \mu \neq \mu_0}$

$Z_{0.975} = 1.96 < 2.844$ So, Z-Value in H_0 Rejected Area (H_1 Accepted area) So, $\mu \neq 8$: Production is NOT under control.

Example: σ unknown & Two Tailed test

A manufacturer claims that the thickness of the spearmint gum it produces is 7.5 one-hundredths of an inch. A quality control specialist regularly checks this claim. On one production run, he took a random sample of *n* = 10 pieces of gum and measured their thickness. He obtained:

The quality control specialist's hypotheses are: H_0: μ = 7.5 vs. H_1: $\mu \neq$ 7.5

7.65	7.60	7.65	7.70	7.55
7.55	7.40	7.40	7.50	7.50

Answer:

<u>Step 01:</u> Defining Hypothesis Testing.

$H_0 : \mu = 7.5 \quad H_1 : \mu \neq 7.5$

<u>Step 02:</u> Test Statistics

Pieces of Gum	Thickness	$(X_i - \bar{X})^2$
1	7.65	0.010
2	7.60	0.002
3	7.65	0.010
4	7.70	0.023
5	7.55	0.000
6	7.55	0.000
7	7.40	0.022
8	7.40	0.022
9	7.50	0.002
10	7.50	0.002
	AVG = 7.55	**SUM =0.095**

In here,

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

$$S^2 = \frac{0.095}{9} = 0.011$$

$$S = 0.103$$

So, the Test Statistics: **T**

Formula: $T = \frac{\bar{x}-\mu}{s/\sqrt{n}}$ $\quad$ $T = \frac{\bar{x}-\mu}{s/\sqrt{n}}$

$$T = \frac{7.55-7.5}{0.103/\sqrt{10}} = 1.5351$$

<u>Step 03:</u> Conclusion

$\boxed{H_0 : \mu = \mu_0}$

$H_1 : \mu \neq \mu_0$

$1.5351 < T_{0.975,9} = 2.262$ So, T-value in H_0 accepted Area (H_1 rejected area) So, $\mu = 7.5$: Production is under control.

So, there is **not enough evidence** to say that the mean thickness is different from 7.5 hundredths of an inch.

Example: σ unknown & One Tailed test (Right Tailed)

An engineer measured the Brinell hardness of 25 pieces of ductile iron that were sub critically annealed. The resulting data were:

170	167	174	179	179
156	163	156	187	156
183	179	174	179	170
156	187	179	183	174
187	167	159	170	179

The engineer hypothesized that the mean Brinell hardness of all such ductile iron pieces is greater than 170. Therefore, he was interested in testing the hypotheses: H_0: μ = 170
H_A: μ > 170

Answer:

<u>Step 01:</u> Defining Hypothesis Testing.

$H_0 : \mu = 170 \quad H_1 : \mu > 170$

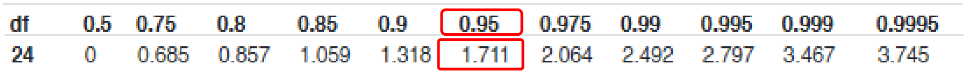
<u>Step 02:</u> Test Statistics

$\sum (X_i - \bar{X})^2$	Average	Standard Deviation	Variance
2552.24	172.52	10.312	106.343

Formula: $T = \frac{\bar{x}-\mu}{s/\sqrt{n}}$

$$T = \frac{172.52-170}{10.312/\sqrt{25}} = 1.2219$$

<u>Step 03:</u> Conclusion



df	0.5	0.75	0.8	0.85	0.9	0.95	0.975	0.99	0.995	0.999	0.9995
24	0	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745

$1.2219 < T_{0.95,24} = 1.711$ So, T-value in H_0 accepted Area (H_1 rejected area) So,

$\boxed{H_0 : \mu \le \mu_0}$

$H_1 : \mu > \mu_0$

$So,\ \mu \le 170$

HYPOTHESIS TEST: DIFFERENCE BETWEEN MEANS

This lesson explains how to conduct a hypothesis test for the difference between two means.

The test procedure, called the **two-sample t-test**, is appropriate when the following conditions are met:

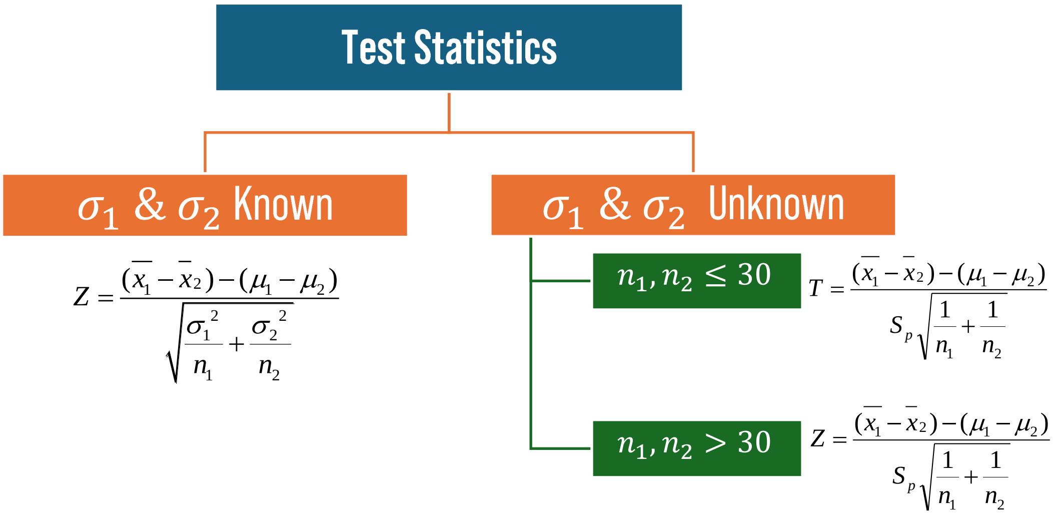
- The sampling method for each sample is simple random sampling.
- The samples are independent.
- Each population is at least 20 times larger than its respective sample.

The sampling distribution is approximately normal, which is generally the case if any of the following **conditions** apply;

- The population distribution is normal.
- The population data are symmetric, unimodal, without outliers.

This approach consists of <u>four steps</u>:

1. state the hypotheses
2. formulate an analysis plan
3. analyze sample data
4. interpret results



Test Statistics

σ_1 & σ_2 Known

$$Z = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

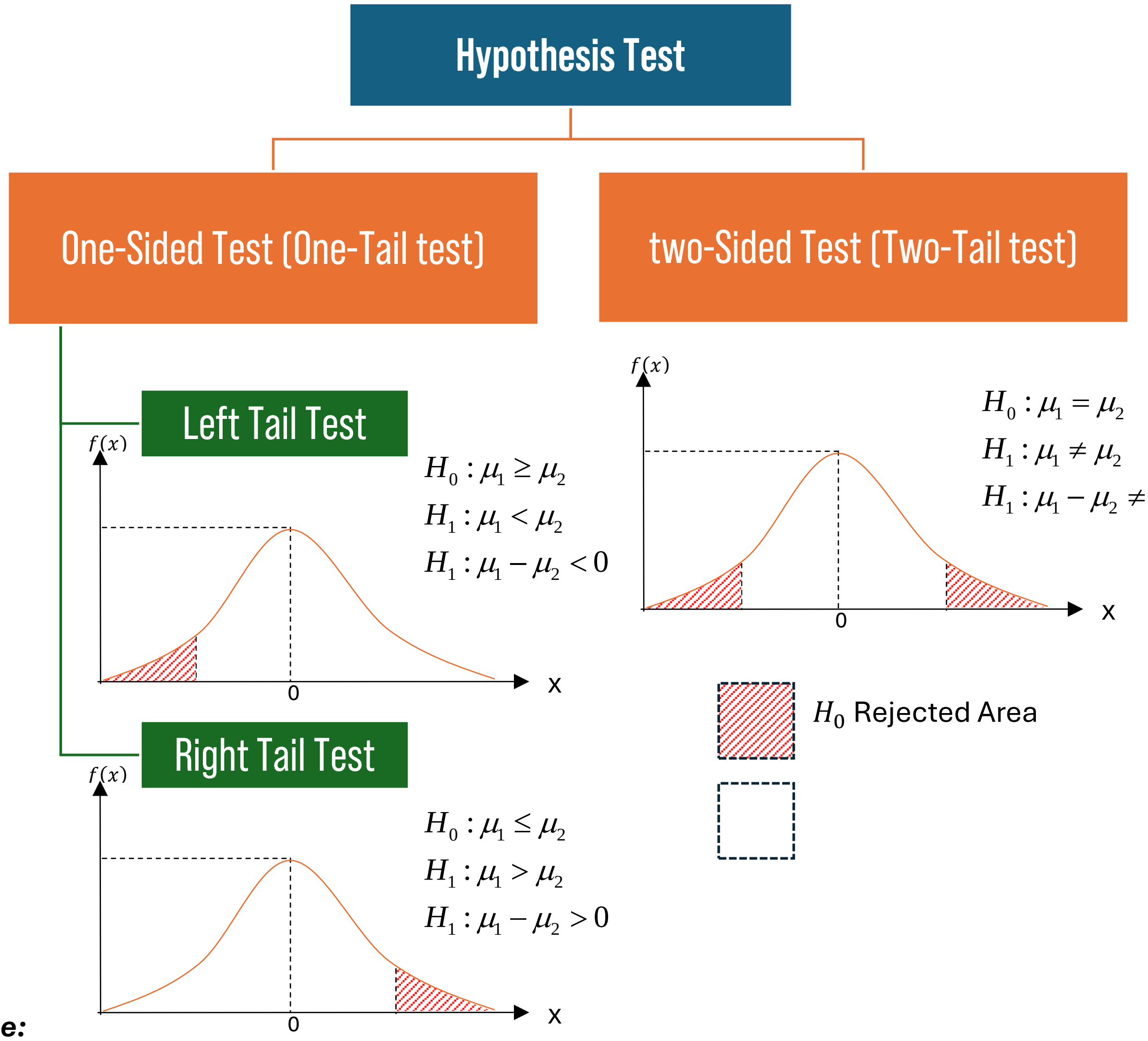
σ_1 & σ_2 Unknown

$n_1, n_2 \le 30$

$$T = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$n_1, n_2 > 30$

$$Z = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$



Hypothesis Test

One-Sided Test (One-Tail test)

two-Sided Test (Two-Tail test)

Left Tail Test

$H_0 : \mu_1 \geq \mu_2$

$H_1 : \mu_1 < \mu_2$

$H_1 : \mu_1 - \mu_2 < 0$

Right Tail Test

$H_0 : \mu_1 \leq \mu_2$

$H_1 : \mu_1 > \mu_2$

$H_1 : \mu_1 - \mu_2 > 0$

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$

$H_1 : \mu_1 - \mu_2 \neq 0$

▨ H_0 Rejected Area

□

Example:

The Acme Company has developed a new battery. The engineer in charge claims that the new battery will operate continuously for at least 7 minutes longer than the old battery.

To test the claim, the company selects a simple random sample of 100 new batteries and 100 old batteries. The old batteries run continuously for 190 minutes with a standard deviation of 20 minutes; the new batteries, 200 minutes with a standard deviation of 40 minutes.

Test the engineer's claim that **the new batteries run at least 7 minutes longer than the old**. Use a **0.05 level of significance**. (Assume that there are no outliers in either sample.)

Answer:

	New	Old
Mean	200 min	190 min
Standard Deviation(S)	40 min	20 min
Sample Size	100	100

$\mu_{new} - \mu_{old} \geq 7$ ➔ $\mu_{new} \geq \mu_{old}$

Sample size(n)>30

Population Standard Deviation (σ): **Unknown**

$$S_p = \sqrt{\frac{S_1{}^2(n_1-1)+S_2{}^2(n_2-1)}{(n_1+n_2-2)}}$$

$$S_p = \sqrt{\frac{20^2\times(100-1)+40^2\times(100-1)}{(200-2)}}$$

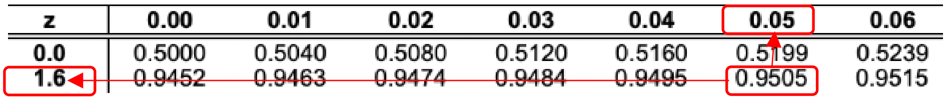
$$S_p = 31.623$$

$$Z = \frac{\left(\overline{X}_{new}-\overline{X}_{old}\right)-\left(\mu_{new}-\mu_{old}\right)}{S_p\sqrt{\frac{1}{n_{new}}+\frac{1}{n_{old}}}}$$

$$Z = \frac{(200-190)-(7)}{31.623\times\sqrt{\frac{1}{100}+\frac{1}{100}}} = 0.6708$$

$\alpha = 0.05$

$1-\alpha = 0.95$



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515

$So,\ Z_{0.95} = 1.65$

$0.6708 < Z_{0.95} = 1.65$

Z-Value from Test Statistics in H_0 Acceptet Area Then Accept Null Hypothesis(H_0)

COMPARING TWO POPULATION MEANS: PAIRED DATA THE PAIRED T- PROCEDURE

In past cases we analyze totally independence difference samples but here we analyze sample with difference situations.

Hypothesis Conditions

$H_0 : \mu_D = \mu_0$

$H_1 : \mu_D \neq \mu_0$

OR

$H_0 : \mu_D = 0$

$H_1 : \mu_D < 0$

OR

$H_0 : \mu_D = 0$

$H_1 : \mu_D > 0$

$\mu_D = Diffrence\ \ between\ \ two\ \ Means$

Test Statistic

Paired t-Interval:

$$\bar{d} \pm t_{\alpha/2} \cdot \frac{S_d}{\sqrt{n}}$$

Equation 11: Test statistics for Paired data

where n denotes the number of pairs or the number of differences

Example:

Using the above example with n = 20 students, the following results were obtained:

Student	Pre-module score	Post-module score
1	18	22
2	21	25
3	16	17
4	22	24
5	19	16
6	24	29
7	17	20
8	21	23
9	23	19
10	18	20
11	14	15
12	16	15
13	16	18
14	19	26
15	18	18
16	20	24
17	12	18
18	22	25
19	15	19
20	17	16

Answer:

Hypothesis Testing

$$H_0 : \mu_{\text{Post}} = \mu_{\text{Pre}}$$

$$H_1 : \mu_{\text{Post}} > \mu_{\text{Pre}}$$

Test Statistics

$$\overline{D} = 2.05$$

$$S_D = 2.837$$

$$T = \frac{\overline{D}}{S_d / \sqrt{n}}$$

$$= \frac{2.05}{2.837 / \sqrt{20}} = 3.23$$

$$\textit{One tail Test So, The Significance Level } (\alpha) = 0.05$$

$$T_{(0.95,19)} = 1.729$$

$$So,\ T_{(0.95,19)} < 3.23$$

T-Value is in H_0 Rejected Area,

So, The Result: $\mathrm{H}_1 : \mu_{Post} > \mu_{Pre}$

ANOVA (ANALYSIS OF VARIANCE)

WHAT IS ANOVA?

ANOVA is a method to **compare 3 or more group means** to see if at least one is significantly different.

WHY USE ANOVA (INSTEAD OF MANY T-TESTS)?

Because doing multiple t-tests increases **error risk**. ANOVA controls that risk.

BASIC IDEA

ANOVA checks if **differences between groups** are **larger than differences within groups**.

WHAT IT DOES

It tests:

H_0 (Null Hypothesis): All group means are equal
H_1 (Alternative Hypothesis): At least one group mean is different

LAYOUT OF THE DESIGN

A diagram that shows the arrangement of treatments (Groups) in experimental units is called as the layout of the design. Consider an experiment with 4 treatments and 5 replicates for each treatment:

Determine the total number of experimental units (n) as the number of treatments (t) and number of replications (r).

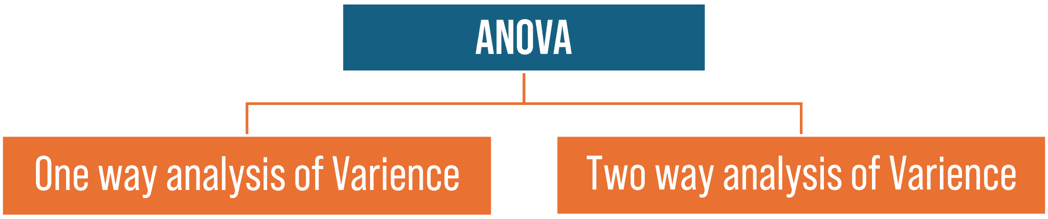
$$n = tr = 4 \times 5 \ = 20$$

We need n, 20 experimental units to carry out this experiment with 4 treatments and 5 replicates.

The 20 units are numbered as follows, Layout of the design by using CRD Experiment (before randomized).

(CRD (*Completely Randomized Design*): This is a Theoretical Ways of Numbering Based on design type.)

Assign the treatments to the experimental units by randomly,



ONE-WAY ANOVA

What is this test for?

One-way ANOVA means we test here only one property here,

Ex: If 3 types of fertilizers introduce for a particular fungal disease, then the result will be focused on only one property which is which fertilizer more effective for that disease.

The one-way ANOVA compares the means between the groups you are interested in and determines whether any of those means are statistically significantly different from each other.

Specifically, it tests the null hypothesis:

$$H_0 = \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$$

$\xrightarrow{Tretment\,(i)}$

1	2	3	.	.	.	s	
y_{11}	y_{21}	y_{31}	.	.	.	y_{s1}	
y_{12}	y_{22}	y_{32}	.	.	.	y_{s2}	
y_{13}	y_{23}	y_{33}	.	.	.	y_{s3}	
.	.	.	.	.	.		
.	.	.	.	.	.		
y_{1n_1}	y_{2n_2}	y_{3n_3}	.	.	.	y_{sn_k}	
$T_{1.}$	$T_{2.}$	$T_{3.}$	.	.	.	$T_{s.}$	$T_{..}$

$$T_{i.} = \sum_{j=1}^{n_i} y_{ij}$$

Equation 12: Treatment Total

$$T_{..} = \sum_{i=1}^{s}\sum_{j=1}^{n_i} y_{ij}$$

Equation 13:Grand Total

$$\overline{T}_{..} = \frac{\sum_{i=1}^{s}\sum_{j=1}^{n_i} y_{ij}}{N} \qquad N = \sum_{i=1}^{k} n_i$$

Equation 14: Grand Total mean

Models for the Data

The statistical model of the CRD for the single factor experiment with t treatments and r replicates, is of the form (Total number of observations, $\boldsymbol{n = tr}$).

Means Model

$y_{ij} = \mu_i + \varepsilon_{ij} \qquad ; i = 1,\ 2,\ \ldots,\ s : j = 1,\ 2,\ \ldots,\ n$

$\mu_i = \mu + \alpha_i : i = 1,\ 2,\ \ldots,\ s$

Effect Model

$y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \qquad ; i = 1,\ 2,\ \ldots,\ s \qquad : j = 1,\ 2,\ \ldots,\ n$

$\boldsymbol{y_{ij}} =$ the jth observation on the ith treatment

$\boldsymbol{\mu}\ =$ grand (overall) mean

$\alpha_i =$ effect of ith treatment

$\boldsymbol{\varepsilon_{ij}} =$ random error

Then the model is,

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \qquad i=1......s \,\&\, j=1.......n_i$$

Equation 15: One-Way ANOVA model

We can replace μ_i by $\mu + \alpha_i$ where μ is the mean of the population and α_i is the effect of i^{th} treatment subject to the **constraint**.

$\alpha_i =$ the **effect** of the ith treatment (how much it differs from the grand mean)

$$\sum_{i=0}^{k} \alpha_i = 0\text{: }\textbf{Contrain}$$

Reason	Why Constraint $\sum_{i=1}^{k} \alpha_i = 0$ is Needed
Centering	Keeps μ as the grand mean
No redundancy	Avoid over-explaining data
Proper Variance split	Allows correct ANOVA calculation

Assumptions and Hypothesis in One-way ANOVA

Assumptions

To apply or perform a one−way ANOVA, certain assumptions (or conditions) need to exist.

If any of the conditions are not satisfied, the results from the use of ANOVA techniques may be unreliable.

The assumptions are:

- **Each sample is an independent random sample**
  a. Errors are normally distributed with mean zero and variance (σ^2),

$$\varepsilon_{ij} \sim N(0, \sigma^2)$$

- **All populations involved follow a normal distribution.**
  a. We shall assume that random variable y_{ij} which are all independent, have the normal distribution with mean μ_i and common variance σ^2**(3rd point)**.

$$y_{ij} \sim N(\mu_i, \sigma^2)$$

- **The sample variances (standard deviation) are equal across responses for the group levels.**

(This can be evaluated by using the following **rule of thumb**: if the largest sample standard deviation divided by the smallest sample standard deviation is not greater than two, then assume that the population variances are equal.)

Hypothesis

In the means model: We are interested to testing the equality of the t treatment means

$$H_o: \mu_1 = \mu_2 = \cdots = \mu_t$$

$$H_A: \mu_i \neq \mu_j \text{ for at least one pair } (i,\ j)$$

In the effects model: We are interested to testing that the treatment effects are zero (There is no treatments/ factor effects)

$$H_o: \alpha_i = 0\ for\ \forall\ i$$

$$H_A: \alpha_i \neq 0 \text{ for at least one } i$$

- The null hypothesis will be that all population means are equal
- The alternative hypothesis is that at least one mean is different.

ANOVA Table for CRD of a Single Factor Experiment: One Way ANOVA Table

Table 1: One-Way ANOVA Table (Used for ANOVA related Calculation)

Source of variance	Sum of Square	Degree of the Freedom	Mean of Square	F-Value
Treatment	SS_{Tr}	$S-1$	$\frac{SS_{Tr}}{S-1} = \alpha$	$\frac{\alpha}{\beta} = f_0$
Error	SS_E	$N-S$	$\frac{SS_E}{N-S} = \beta$	
Totals	SS_T	$N-1$		

Total Sum Square (SS_T)

$$SS_T = \sum_{i=1}^{s}\sum_{j=1}^{n_i}(y_{ij} - \overline{T}_{..})^2$$

$$\boxed{SS_T = \sum_{i=1}^{s}\sum_{j=1}^{n_i} y_{ij}^2 - \frac{T_{..}^2}{N}}$$

Equation 16: Equation for Total Sum Square (SS$_T$)

Total Variance due to treatment (SS_{Tr})

$$SS_{Tr} = \sum_{i=1}^{s}\sum_{j=1}^{n_i}(\overline{T}_{i.} - \overline{T}_{..})^2$$

$$\boxed{SS_{Tr} = \sum_{i=1}^{s}\frac{T_{i.}^2}{n_i} - \frac{T_{..}^2}{N}}$$

Equation 17: Equation for Total variance due to treatment (SS$_{Tr}$)

$\boxed{SS_T = SS_{Tr} + SS_E}$ $\boxed{SS_E - Error\ sum of\ square}$

To know about $\overline{T}..$, $T..^2$, n_i, N, y_{ij} Got to Page 37

Example

Imagine that you manufacture paper bags and that you want to improve the tensile strength of the bag.

You suspect that changing the concentration of hardwood in the bag will change the tensile strength. You measure the tensile strength in pounds per square inch (PSI).

you decide to test this at 5%, 10%, 15% and 20% hardwood concentration levels. These "levels" are also called "treatments."

Hardwood Concentration (%)				Grand Total
A (5%)	B (10%)	C (15%)	D (20%)	
12	37	15	27	
7	12	18	25	
15	17	10	30	
11	13	19	15	
10	18	16	13	
	19	18		
	15			
$T_1 = 55$	$T_2 = 131$	$T_3 = 96$	$T_4 = 110$	T... = 392

Answer:

Defining the hypothesis:

H_o: $\mu_1 = \mu_2 = \mu_3 = \mu_4$

H_A: $\mu_i \neq \mu_j$ $for\ at\ least\ one\ pair\ (i,\ j)$

Defining Model: One-Way ANOVA Model

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \qquad {\scriptstyle i=1......s\ \&\ j=1.......n_i}$$

Assumption:

- Data sample selected in a random manner

  $\varepsilon_{ij} \sim N(0,\sigma^2)$
- The variance of the sample is same

Constrain:

$$\sum_{i=1}^{k} \alpha_i = 0$$

Calculating:

$$SS_{Tr} = \sum_{i=1}^{s} \frac{T_{i.}^2}{n_i} - \frac{T_{..}^2}{N}$$

$$= \frac{T_1^2}{n_2} + \frac{T_2^2}{n_2} + \frac{T_3^2}{n_3} + \frac{T_4^2}{n_4} - \frac{T_{..}^2}{N}$$

$$= \frac{55^2}{5} + \frac{131^2}{7} + \frac{96^2}{6} + \frac{110^2}{5} - \frac{392}{23}$$

$$SS_{Tr} = 331.528$$

$$SS_T = \sum_{i=1}^{s}\sum_{j=1}^{n_i} y_{ij}{}^2 - \frac{T_{..}^{\ 2}}{N}$$

$$SS_T = 12^2 + 7^2 + 15^2 + 11^2 + 10^2 + 37^2 + ...+ 30^2 + 15^2 + 13^2 - \frac{392}{23}$$

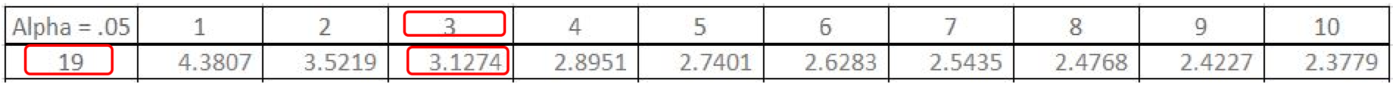
$$SS_T = 1076.975$$

$$SS_E = SS_T - SS_{Tr}$$

$$SS_E = 1076.975 - 331.528 = 745.429$$

Source of variance	Sum of Square	Degree of the Freedom	Mean of Square	F-Value
Treatment	SS_{Tr} =331.528	$S-1$=3	$\frac{SS_{Tr}}{S-1} = \alpha$ =110.509	$\frac{\alpha}{\beta} = f_0$ $f_0 = 2.817$
Error	SS_E =745.429	$N-S$ =19	$\frac{SS_E}{N-S} = \beta$ =39.233	
Totals	SS_T =1076.957	$N-1$=22		

Making Decision



Alpha = .05	1	2	3	4	5	6	7	8	9	10
19	4.3807	3.5219	3.1274	2.8951	2.7401	2.6283	2.5435	2.4768	2.4227	2.3779

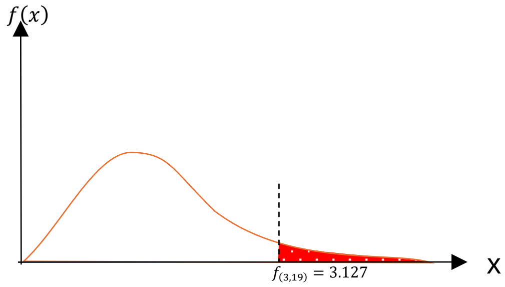
Figure 9:F-Distribution Table

So, here significant level: $\alpha = 0.05$

$$f(2.817) < f_\alpha(3.1274)$$

So, the value 2.817 is H_0 Accepted area... So,

$$H_o\!: \mu_1 = \mu_2 = \mu_3 = \mu_4$$



TWO-WAY ANOVA

In here we consider two factors

As an example, let's assume we're planting corn. The

- type of seed
- type of fertilizer

are the two factors we're considering in this example.

- This example has 15 treatment groups.
- The data that actually appears in the table as follows.

Seed / Fertilizer	Fert I	Fert II	Fert III	Fert IV	Fert V
Seed A-402	106, 110	95, 100	94, 107	103, 104	100, 102
Seed B-894	110, 112	98, 99	100, 101	108, 112	105, 107
Seed C-952	94, 97	86, 87	98, 99	99, 101	94, 98

$\downarrow Block(j)$ $\xrightarrow{Tretment\,(i)}$

	1	2	3	.	.	.	S	
1	y_{11}	y_{21}	y_{31}	.	.	.	y_{s1}	$T_{.1}$
2	y_{12}	y_{22}	y_{32}	.	.	.	y_{s2}	$T_{.2}$
3	y_{13}	y_{23}	y_{33}	.	.	.	y_{s3}	$T_{.3}$
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
b	y_{1b}	y_{2b}	y_{3b}	.	.	.	y_{s3}	$T_{.b}$
	$T_{1.}$	$T_{2.}$	$T_{3.}$	.	.	.	$T_{s.}$	$T_{..}$

We shall assume that random variable y_{ij} which are all independent, have the normal distribution with mean and common variance σ^2.

$$x \sim N(\mu_i, \sigma^2)$$

Model for the given observations.

$$y_{ij} = \mu_i + \varepsilon_{ij} \quad {}_{i=1......s\,\&\,j=1.......n_i} \qquad\qquad \varepsilon_{ij} \quad -random\ \ error$$

We can replace μ_i by $\mu + \alpha_i + \beta_j$ where μ is the mean of the population and α_i is the effect of i^{th} treatment and β_j is the j^{th} block effect.

Then the model is:

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \quad {}_{i=1......s\,\&\,j=1.......b_i}$$

Equation 18: Two-Way ANOVA Model

Assumptions

- The populations from which the samples were obtained must be normally or approximately normally distributed.
- The samples must be independent.
- The variances of the populations must be equal.
- The groups must have the same sample size

$$Model\ : y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \quad {}_{i=1......s\,\&\,j=1.......b}$$

$$Assumption$$

$$\varepsilon_{ij} \sim N(0, \sigma^2)$$

$$\varepsilon_{ij}\ are\, independent$$

$$Constrain$$

$$\sum_{i=1}^{s} \alpha_i = 0 \ \sum_{j=1}^{b} \beta_j = 0$$

Hypotheses

There are three sets of hypotheses with the two-way ANOVA. The null hypotheses for each of the sets are given below.

1. The population means of the first factor are equal. This is like the one-way ANOVA for the row factor.
2. The population means of the second factor are equal. This is like the one-way ANOVA for the column factor.
3. There is no interaction between the two factors.

$$
\begin{aligned}
&H_0 : \mu_{1.} = \mu_{2.} = \mu_{3.} = ... = \mu_{s,} \; (Tretment\, mean\, are\, same)\\
&H_1 : \mu_{1.} \neq \mu_{2.} \neq \mu_{3.} \neq ...... \neq \mu_{s.} \; (Tretment\, mean\, are\, different)\\
&H_0^{/} : \mu_{.1} = \mu_{.2} = \mu_{.3} = ... = \mu_{.b} \; (Block\, mean\, are\, same)\\
&H_1^{/} : \mu_{.b} \neq \mu_{.b} \neq \mu_{.b} \neq ...... \neq \mu_{.b} \; (Tretment\, mean\, are\, different)
\end{aligned}
$$

Table 2: Two-Way ANOVA Table (Used for ANOVA related Calculation)

Source of variance	Sum of Square	Degree of the Freedom	Mean of Square	F-Value
Treatment	$\boxed{SS_{Tr} = \sum_{i=1}^{s} \frac{T_{i.}^{\,2}}{B} - \frac{T_{..}^{\,2}}{BS}}$	$S-1$	$\frac{SS_{Tr}}{S-1} = MSTr$	$\frac{MSTr}{MSe} = f_0$
Block	$\boxed{SS_B = \sum_{j=1}^{b} \frac{T_{.j}^{\,2}}{S} - \frac{T_{..}^{\,2}}{BS}}$	$B-1$	$\frac{SS_B}{B-1} = MSBl$	$\frac{MSBl}{MSe} = f_1$
Error	$\boxed{SS_E = SS_T - \left(SS_{Tr} + SS_B\right)}$	$(S-1)(B-1)$	$\frac{SS_B}{(S-1)(B-1)} = MSe$	
Totals	$\boxed{SS_T = \sum_{i=1}^{s}\sum_{j=1}^{b} y_{ij}^{\,2} - \frac{T_{..}^{\,2}}{N}}$	$N-1$		

Total Sum Square (SS_T)

$$SS_T = \sum_{i=1}^{s}\sum_{j=1}^{b} (y_{ij} - \overline{T}_{..})^2$$

$$\boxed{SS_T = \sum_{i=1}^{s}\sum_{j=1}^{b} y_{ij}^{\,2} - \frac{T_{..}^{\,2}}{N}}$$

Equation 19: Equation for Total Sum Square (SS_T)

Total Variance due to treatment (SS_{Tr})

$$SS_{Tr} = \sum_{i=1}^{s}\sum_{j=1}^{b} (\overline{T}_{i.} - \overline{T}_{..})^2$$

$$\boxed{SS_{Tr} = \sum_{i=1}^{s} \frac{T_{i.}^{\,2}}{B} - \frac{T_{..}^{\,2}}{BS}}$$

Equation 20: Equation for Total variance due to treatment (SS_{Tr})

Total Variance due to block (SS_B)

$$SS_B = \sum_{i=1}^{s}\sum_{j=1}^{b}(\overline{T}_{.j} - \overline{T}_{..})^2$$

$$\boxed{SS_B = \sum_{j=1}^{b}\frac{T_{.j}^{\ 2}}{S} - \frac{T_{..}^{\ 2}}{BS}}$$

$\boxed{SS_T = SS_{Tr} + SS_B + SS_E}$ $\boxed{SS_E - Error\ sum of\ square}$

To know about $\overline{T}..$, $T..^2$, n_i, N, y_{ij} Got to Page 37

Example:

The government Agricultural experiment station conducted an experiment to compare the effects of four different fertilizers A, B, C and D on the yield of a cane crop.

Researchers divided four blocks of soil in to four plots of equal size and shape and assigned the fertilizers to the plots at random, in such a way that each fertilizer was applied once in each block. The yield of cane is given below.

Block	Fertilizer			
	A	B	C	D
1	50	45	44	40
2	55	42	50	46
3	45	43	35	35
4	56	33	45	36

Do these data provide sufficient evidence to indicate a fertilizer effect and soil effect at the 0.05 level of significance?

Answer:

Defining the hypothesis:

$$H_0 : \mu_{1.} = \mu_{2.} = \mu_{3.} = ... = \mu_{s.}\ (Tretment\ mean\ are\ same)$$
$$H_1 : \mu_{1.} \neq \mu_{2.} \neq \mu_{3.} \neq ...... \neq \mu_{s.}\ (Tretment\ mean\ are\ different)$$
$$H_0^{\ /} : \mu_{.1} = \mu_{.2} = \mu_{.3} = ... = \mu_{.b}\ (Block\ mean\ are\ same)$$
$$H_1^{\ /} : \mu_{.1} \neq \mu_{.1} \neq \mu_{.1} \neq ...... \neq \mu_{.b}\ (Block\ mean\ are\ different)$$

Defining Model: Two-Way ANOVA Model

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \qquad {\scriptstyle i=1......s\ \&\ j=1.......b_i}$$

Assumption:

- $\varepsilon_{ij} \sim N(0, \sigma^2)$
- $\varepsilon_{ij}\ are\ \ independent$

Constrain:

$$\sum_{i=1}^{s}\alpha_i = 0 \sum_{j=1}^{b}\beta_j = 0$$

Calculating:

Ti	179	193	158	170
Tj	206	163	174	157

$$SS_{Tr} = \sum_{i=1}^{s}\frac{T_{i.}^{\,2}}{B} - \frac{T_{..}^{\,2}}{BS}$$

$$= \frac{179^2 + 193^2 + 158^2 + 170^2}{4} - \frac{700^2}{4\times 4}$$

$$SS_{Tr} = 163.5$$

$$SS_B = \sum_{j=1}^{b}\frac{T_{.j}^{\,2}}{S} - \frac{T_{..}^{\,2}}{BS}$$

$$= \frac{206^2 + 163^2 + 174^2 + 157^2}{4} - \frac{700^2}{4\times 4}$$

$$SS_B = 357.5$$

$$SS_T = \sum_{i=1}^{s}\sum_{j=1}^{b} y_{ij}^{\,2} - \frac{T_{..}^{\,2}}{N}$$

$$SS_T = 50^2 + 55^2 + 45^2 + 56^2 + 45^2 + 42^2 + \ldots\ldots + 46^2 + 35^2 + 36^2 - \frac{700^2}{16}$$

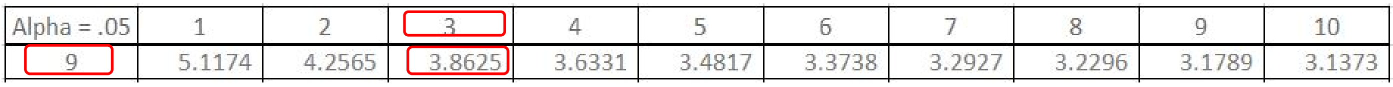
$$SS_T = 711$$

$$SS_E = SS_T - \left(SS_{Tr} + SS_B\right)$$

$$SS_E = 711 - \left(163.5 + 357.5\right) = 190$$

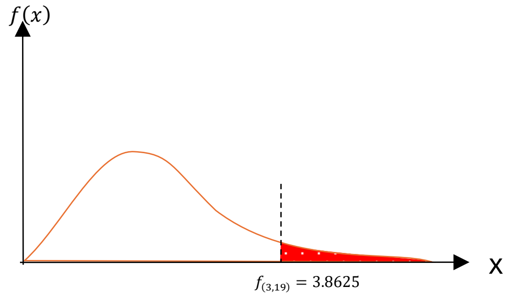
Source of variance	Sum of Square	Degree of the Freedom	Mean of Square	F-Value
Treatment	$SS_{Tr} = 163.5$	$S-1=3$	$\frac{SS_{Tr}}{S-1} = MSTr = 54.5$	$\frac{MSTr}{MSe} = f_0 = 2.582$
Block	$SS_B = 357.5$	$B-1=3$	$\frac{SS_B}{B-1} = MSBl = 119.167$	$\frac{MSBl}{MSe} = f_1 = 5.645$
Error	$SS_E = 190$	$(S-1)(B-1) = 9$	$\frac{SS_B}{(S-1)(B-1)} = MSe = 21.111$	
Totals	$SS_T = 711$	$N-1=15$		

Making Decision



Alpha = .05	1	2	3	4	5	6	7	8	9	10
9	5.1174	4.2565	3.8625	3.6331	3.4817	3.3738	3.2927	3.2296	3.1789	3.1373

So, here significant level: $\alpha = 0.05$



$f_0(2.582) < f_\alpha(3.8625)$

So, the value 2.817 is H₀ Accepted area... So,

$H_o\colon\ \mu_1 = \mu_2 = \mu_3 = \mu_4$(Soil has not Effect Exist)

$f_1(5.645) > f_\alpha(3.8625)$

So, the value 2.817 is H₀' Rejected area... So,

$H_1\colon\ \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$(Fertilizers have Effect!)

ANALYSIS OF VARIANCE OF A RANDOMIZED BLOCK DESIGN WITH MISSING OBSERVATIONS

	Verities of Potatoes				
		V	W	X	$T_{.j}$
Fertilizer Brand	A	55	72	47	174
	B	64	66	53	183
	C	Y	57	74	131
	D	59	57	58	174
	$T_{i.}$	178	252	232	x.. =662

$$x_{i.j} = \frac{rx_{i.} + cx_{.j} - x..}{(r-1)(c-1)}$$

Equation 21:Finding Missing Value

r = # of rows (blocking factor)

c = # of columns (treatment factor).

x_{ij} = value of the cell in the i^{th} row and j^{th} column

$x_{i\cdot}$ = the sum of the values in the i^{th} row,

$x_{.j}$ = the sum of the values in the j^{th} column

$x..$ = the sum of all the values.

$$x_{i.j} = \frac{rx_{i.} + cx_{.j} - x..}{(r-1)(c-1)}$$

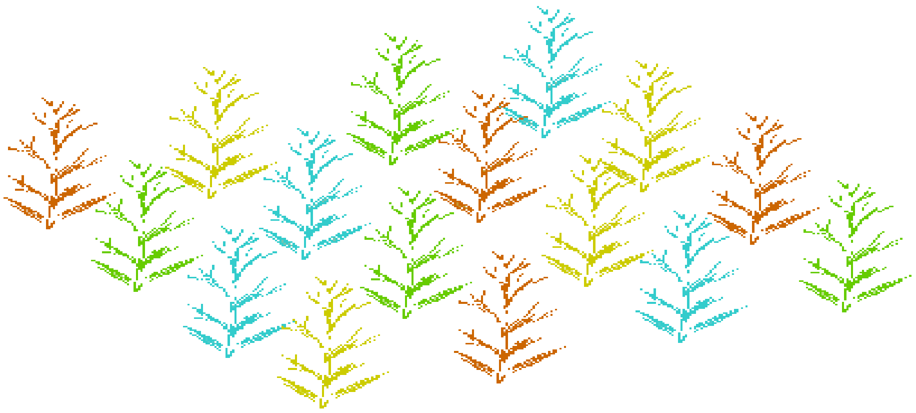
$$x_{i.j} = \frac{4\times178 + 4\times131 - 662}{3\times3}$$

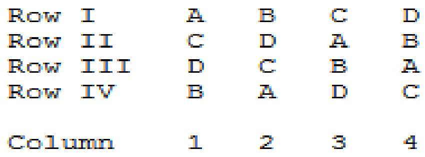
$$x_{i.j} = 63.778$$

THE LATIN SQUARE DESIGN

This experimental Design used to check three factors in an experiment at a time.

SAMPLE LAYOUT:





```
Row I       A    B    C    D
Row II      C    D    A    B
Row III     D    C    B    A
Row IV      B    A    D    C

Column      1    2    3    4
```

- Different colors represent different treatments.
- There are **4 treatments** (A-D) assigned to **4 rows** (I-IV) and **4 columns** (1-4).

DEFINITION

Treatments are assigned **randomly** within each **row** and **column**, but **each treatment appears only once per row and once per column**.

STRUCTURE

- Same number of rows, columns, and treatments.
- Common layout: a square (e.g., 4 treatments → 4 rows × 4 columns (Square shape: p × p)).
- Each treatment appears exactly once in each row and each column

	Time				
Days		1	2	3	4
	1	P1: 11	P2: 13	P3: 27	P4: 34
	2	P2: 16	P3: 33	P4: 28	P3: 34
	3	P3: 31	P4:23	P1: 48	P2: 29
	4	P4: 61	P1: 58	P2: 52	P1: 64

Treatments

Column

Column

Parametric model

$$Y = X_{ijk} = \mu + \alpha_i + \beta_i + \gamma_k + \in_{ijk}$$

Where each of i, j and k reanges from 1 to t.

α_i : Effect due to treatment i.

β_i : Effect due to row j.

γ_k : Effect due to column k.

Constrains

There is a three constrain.

$$\sum_{i=1}^{k} \alpha_i = 0 \qquad \sum_{i=1}^{k} \beta_j = 0 \qquad \sum_{i=1}^{k} \gamma_k = 0$$

α_i : Effect due to treatment i.

β_i : Effect due to row j.

γ_k : Effect due to column k.

Assumption

Homogeneity of Variance (Constant Error Variance)

- The variance of experimental errors (ϵ_{ijk}) is assumed to be constant across all treatments, rows, and columns.
- Explanation: This means that the spread of observations around their means should be roughly equal.

Independence of Errors

- The errors (ϵ_{ijk}) are assumed to be independent of each other.
- Purpose: This ensures that there is no systematic bias affecting certain groups more than others.

Normality of Errors

- The errors (ϵ_{ijk}) are normally distributed.
- Importance: This assumption is crucial for valid ANOVA and F-tests used in LSD (Least Significant Difference) analysis.

No Missing Observations

- A complete Latin square requires all treatments to appear exactly once per row and per column.

Hypothesis

$$H_0 : \mu_{1.} = \mu_{2.} = \mu_{3.} = ... = \mu_{s,}\ (Tretment\ mean\ are\ same)$$

$$H_1 : \mu_{1.} \neq \mu_{2.} \neq \mu_{3.} \neq ...... \neq \mu_{s.}\ (Tretment\ mean\ are\ different)$$

$$H_0^{/} : \mu_{.1} = \mu_{.2} = \mu_{.3} = ... = \mu_{.b}\ (Row\ mean\ are\ same)$$

$$H_1^{/} : \mu_{.1} \neq \mu_{.1} \neq \mu_{.1} \neq ...... \neq \mu_{.b}\ (Row\ mean\ are\ different)$$

$$H_0^{//} : \mu_{.1} = \mu_{.2} = \mu_{.3} = ... = \mu_{.b}\ (Column\ mean\ are\ same)$$

$$H_1^{//} : \mu_{.1} \neq \mu_{.1} \neq \mu_{.1} \neq ...... \neq \mu_{.b}\ (Column\ mean\ are\ different)$$

Table 3:Latin Square Design: ANOVA Table

Source of variance	Sum of Square	Degree of the Freedom	Mean of Square	F-Value
Treatment	$SS_{Tr} = \sum_{j=1}^{p} \frac{T_{.j}^{\ 2}}{P} - \frac{y_{..}^{\ 2}}{P^2}$	$P-1$	$\frac{SS_{Tr}}{P-1} = MSTr$	$f_0 = \frac{MSTr}{MSe}$
Row	$SS_R = \sum_{i=1}^{p} \frac{R_i^{\ 2}}{P} - \frac{y_{...}^{\ 2}}{P^2}$	$P-1$	$\frac{SS_R}{P-1} = MSR$	$f_R = \frac{MSR}{MSe}$
Column	$SS_C = \sum_{k=1}^{p} \frac{C_k^{\ 2}}{P} - \frac{y_{...}^{\ 2}}{P^2}$	$P-1$	$\frac{SS_C}{P-1} = MSC$	$f_C = \frac{MSC}{MSe}$
Error	$SS_E = SS_T - \left(SS_{Tr} + SS_C + SS_R\right)$	$(P-1)(P-2)$	$\frac{SS_E}{(P-1)(P-2)} = MSe$	
Totals	$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{b} y_{ijk}^2 - \frac{y_{...}^2}{p^2}$	$P^2 - 1$		

Total Sum Square (SS_T)

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{b} (y_{ijk} - \overline{y}_{..})^2$$

$$\boxed{SS_T = \sum_{i=1}^{a}\sum_{j=1}^{b} y_{ijk}^2 - \frac{y_{...}^2}{p^2}}$$

Equation 22: Equation for Total Sum Square (SS$_T$)

Total Variance due to treatment (SS_{Tr})

$$SS_{Tr} = \sum_{j=1}^{p} p\left(\overline{y}_{.j} - \overline{y}_{..}\right)^2$$

$$\boxed{SS_{Tr} = \sum_{j=1}^{p} \frac{T_{.j}^{\ 2}}{P} - \frac{y_{..}^{\ 2}}{P^2}}$$

Equation 23: Equation for Total variance due to treatment (SS$_{Tr}$)

Total Variance due to row (SS_R)

$$SS_R = \sum_{i=1}^{p} p\left(\overline{y_{i..}} - \overline{y_{..}}\right)^2$$

$$\boxed{SS_R = \sum_{i=1}^{p} \frac{R_i^2}{P} - \frac{y_{...}^2}{P^2}}$$

Total Variance due to Column (SS_C)

$$SS_C = \sum_{k=1}^{p} p\left(\overline{y_{..k}} - \overline{y_{...}}\right)^2$$

$$\boxed{SS_C = \sum_{k=1}^{p} \frac{C_k^2}{P} - \frac{y_{...}^2}{P^2}}$$

$$\boxed{SS_E = SS_T - \left(SS_{Tr} + SS_C + SS_R\right)} \quad \boxed{SS_E - Error\ sum of\ square}$$

Example

A researcher aims to study the effect of four different fertilizers (**A, B, C, and D**) on plant growth. However, plant growth is also influenced by two nuisance factors: **soil type** and **irrigation level**.

To control these nuisance factors, the researcher employs a **4 × 4 Latin square design**, where:

- **Treatments (T):** Fertilizers (**A, B, C, D**)
- **Rows (R):** Different soil types
- **Columns (C):** Different irrigation levels

Latin Square Layout

Soil Type \ Irrigation	1	2	3	4
1	A	B	C	D
2	B	C	D	A
3	C	D	A	B
4	D	A	B	C

Let's assume the measured yield (in kg per plant) is:

Soil Type \ Irrigation	1	2	3	4
1	15	18	14	19
2	17	16	20	15
3	16	19	18	14
4	18	17	15	20

Answer:

Row Total	66	66	66	66
Column Total	66	70	67	68

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{b} y_{ijk}^2 - \frac{y_{...}^2}{p^2}$$

$$= 15^2 + 18^2 + 14^2 + 19^2 + 17^2 + \ldots + 18^2 + 17^2 + 15^2 + 20^2 - \frac{271^2}{4^2}$$

$$SS_T = 60.938$$

$$SS_{Tr} = \sum_{j=1}^{p} \frac{T_{.j}^{\ 2}}{P} - \frac{y_{..}^{\ 2}}{P^2}$$

$$= \frac{65^2 + 64^2 + 66^2 + 76^2}{4} - \frac{271^2}{4^2}$$

$$SS_{Tr} = 23.188$$

$$SS_R = \sum_{i=1}^{p} \frac{R_i^{\ 2}}{P} - \frac{y_{...}^{\ 2}}{P^2}$$

$$= \frac{66^2 + 68^2 + 67^2 + 70^2}{4} - \frac{271^2}{4^2}$$

$$SS_R = 2.1875$$

$$SS_C = \sum_{k=1}^{p} \frac{C_k^{\ 2}}{P} - \frac{y_{...}^{\ 2}}{P^2}$$

$$= \frac{66^2 + 70^2 + 67^2 + 68^2}{4} - \frac{271^2}{4^2}$$

$$SS_C = 2.1875$$

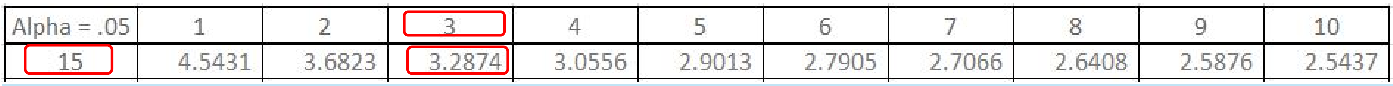
$$SS_E = SS_T - \left(SS_{Tr} + SS_C + SS_R\right)$$

$$SS_E = 60.938 - \left(23.188 + 2.1875 + 2.1875\right) = 33.375$$

Making Decision

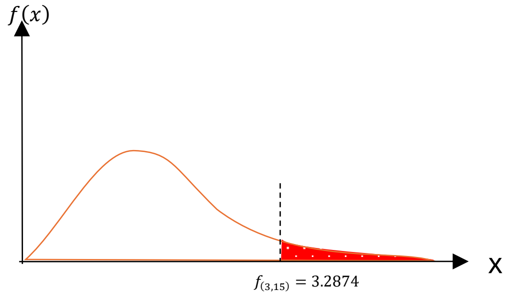
So, here significant level: $\alpha = 0.05$

Degree of freedom (P^2-1), (P-1)➔ (15), (3)



Alpha = .05	1	2	3	4	5	6	7	8	9	10
15	4.5431	3.6823	3.2874	3.0556	2.9013	2.7905	2.7066	2.6408	2.5876	2.5437

Mean of Square	F-Value
$MSTr = 7.729$	$f_0 = 1.39$
$MSR = 0.729$	$f_R = 0.131$
$MSC = 0.729$	$f_C = 0.131$
$MSe = 5.563$	



$f_0(1.39) < f_\alpha(3.2874)$

So, the value 1.39 is H_0 Accepted area… So,

H_o: $\mu_1 = \mu_2 = \mu_3 = \mu_4$

$$f_R(0.131) < f_\alpha(3.2874)$$

So, the value 0.131 is H₀ Accepted area... So,

$$H_o^{/}: \mu_1 = \mu_2 = \mu_3 = \mu_4$$

$$f_C(0.131) < f_\alpha(3.2874)$$

So, the value 0.131 is H₀ Accepted area... So,

$$H_o^{//}: \mu_1 = \mu_2 = \mu_3 = \mu_4$$

LATIN SQUARE DESIGN WITH MISSING OBSERVATIONS

We can find the missing value here as well

$$Y_{i.j.k} = \frac{P\left(R_i + C_j + T_k\right) - 2G}{(P-1)(P-2)}$$

$\hat{Y}_{ijk}$: Estimated value of the missing observation in row i, column j, with treatment k.

P : Number of rows, columns, and treatments in the Latin Square.

R_i : Sum of observed values in the i^{th} row (excluding the missing value).

C_j : Sum of observed values in the j^{th} column (excluding the missing value).

T_k : Sum of observed values in the k^{th} Treatment (excluding the missing value).

G : Grand total of all observed values in the Latin Square.