



CHRIST
(DEEMED TO BE UNIVERSITY)
BANGALORE · INDIA

Applied Statistics Using R

Unit-3

MISSION

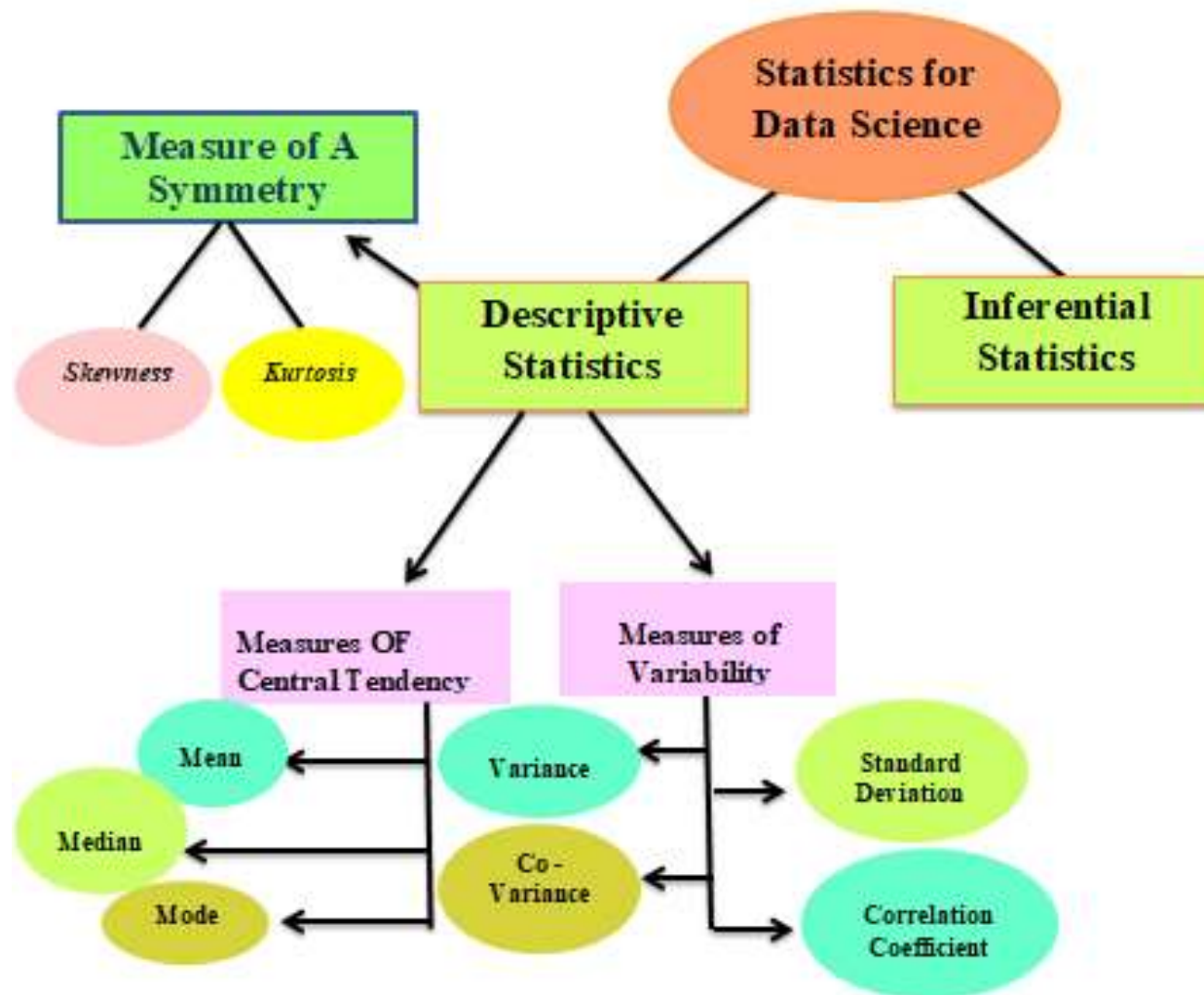
CHRIST is a nurturing ground for an individual's holistic development to make effective contribution to the society in a dynamic environment

VISION

Excellence and Service

CORE VALUES

Faith in God | Moral Uprightness
Love of Fellow Beings
Social Responsibility | Pursuit of Excellence



CENTRAL TENDENCY

1. **Mean** = Sum of scores divided by the number of scores (often referred to as the statistical average)

Pronounced "x-bar" $\bar{X} = \frac{\sum X}{N}$ Capital Sigma for "Sum of"

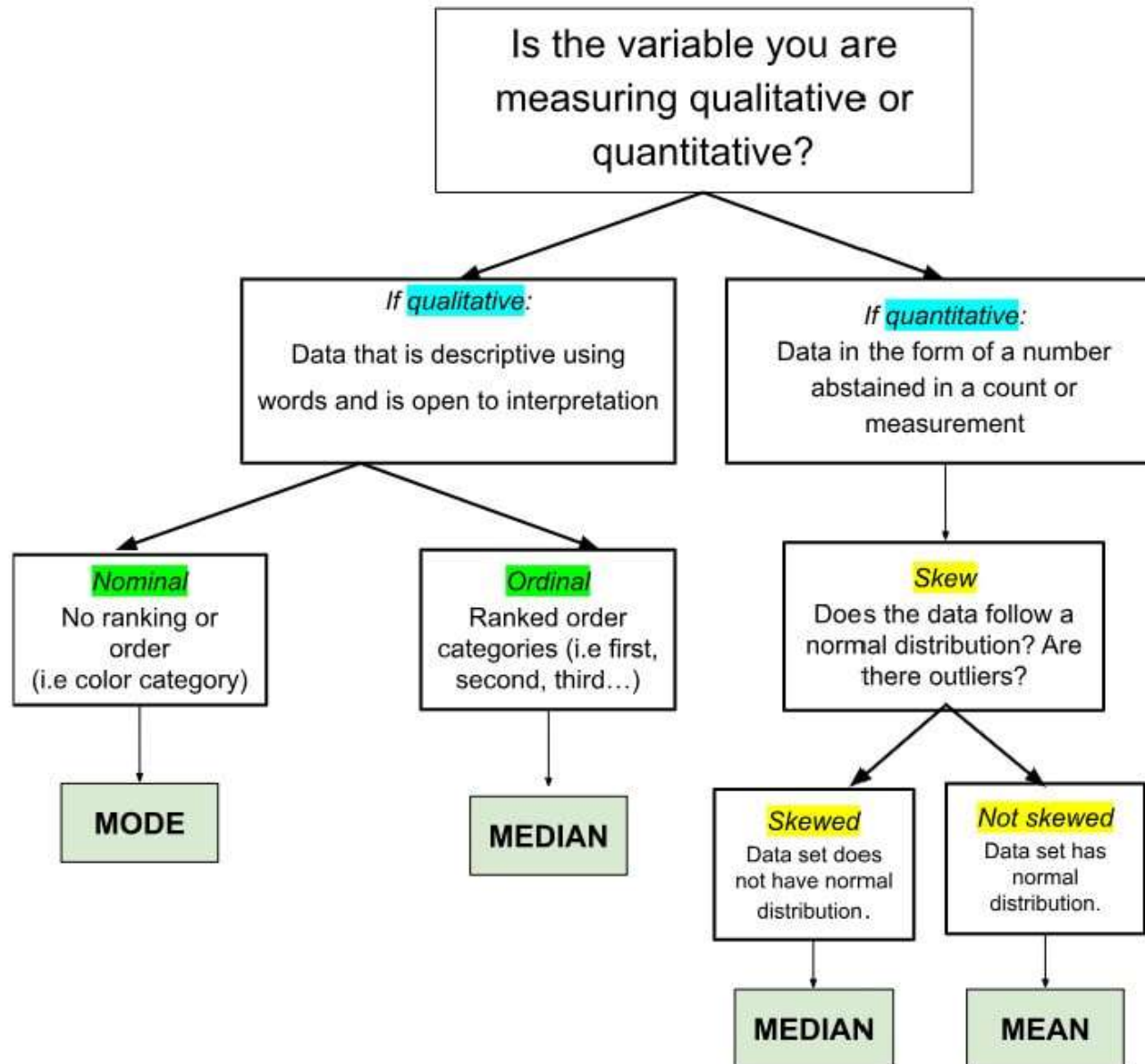
N represents the number of scores \bar{X} "X" represents each score

2. **Median** = Middle Most Number

$$M_d$$

3. **Mode** = Most Frequently Occurring Number

$$M_o$$



Measures Variability

- Range
- Standard Deviation
- Variance

B18				=SQRT(D16/(10-1))
	A			
4		A	B	C = B^2
5	No.	Returns	Return - Mean	(Return - Mean)^2
6	1	72	12	144
7	2	45	-15	225
8	3	58	-2	4
9	4	84	24	576
10	5	60	0	0
11	6	10	-50	2500
12	7	91	31	961
13	8	65	5	25
14	9	55	-5	25
15	10	60	0	0
16	Total	600		4460
17	Mean	60		
18	Standard Deviation	22.26		
19				

Example

Find the **standard deviation** and **variance**

x	$x - \bar{x}$	$(x - \bar{x})^2$
30	4	16
26	0	0
<u>22</u>	-4	16
78		<u>32</u>

Mean = 26

Sum = 0

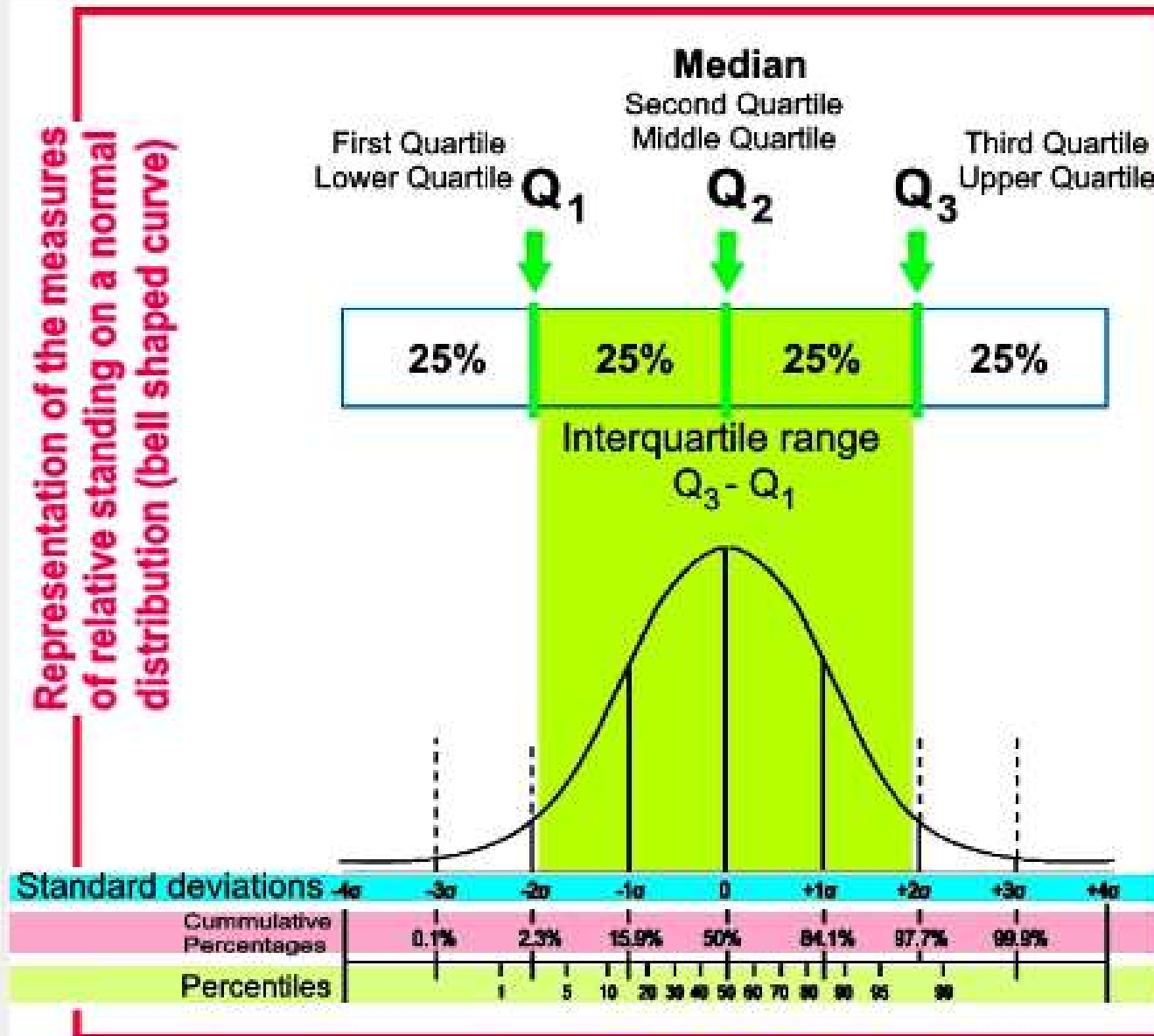
The **variance**

$$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1} = 32 \div 2 = 16$$

The **standard deviation**

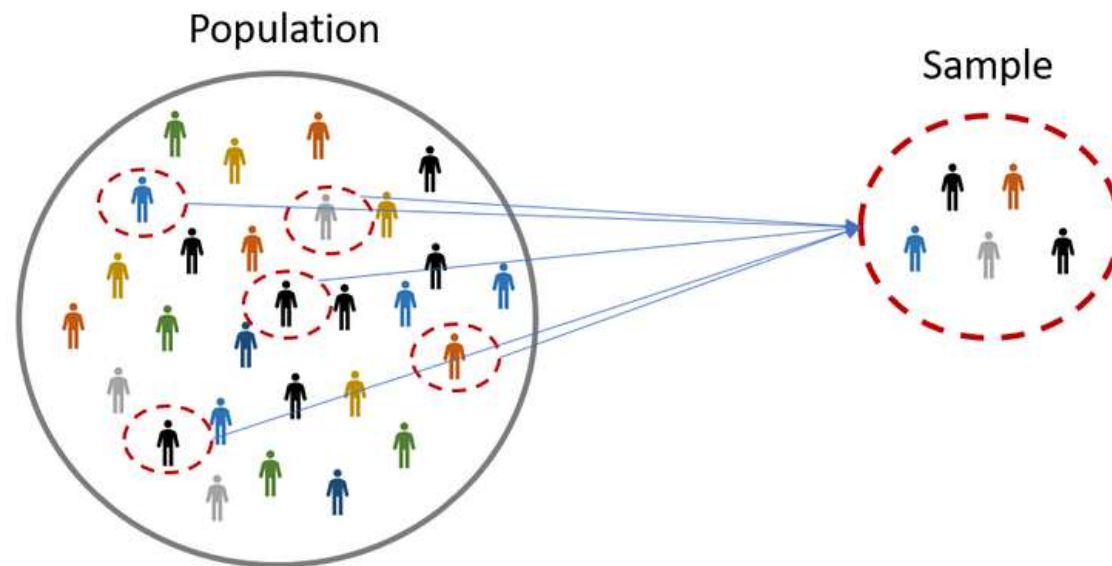
$$s = \sqrt{16} = 4$$

Measures of Relative Position: Quartiles, Percentiles.



Population Vs Sample

- **Population** : The Population is the Entire group that you are taking for analysis or prediction.
- **Sample** : Sample is the Subset of the Population(i.e. Taking random samples from the population). The size of the sample is always less than the total size of the population.



Variance, Standard Deviation, Range

Dispersion	Population	Dispersion	Sample
Variance	$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$	Variance	$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$
Standard Deviation	$\sigma = \sqrt{\frac{\sum (X - \mu)^2}{N}}$	Standard Deviation	$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$
Range	Max – Min	Range	Max – Min

➤ μ = Mean

➤ X = Random Variable

➤ N = no. of data types

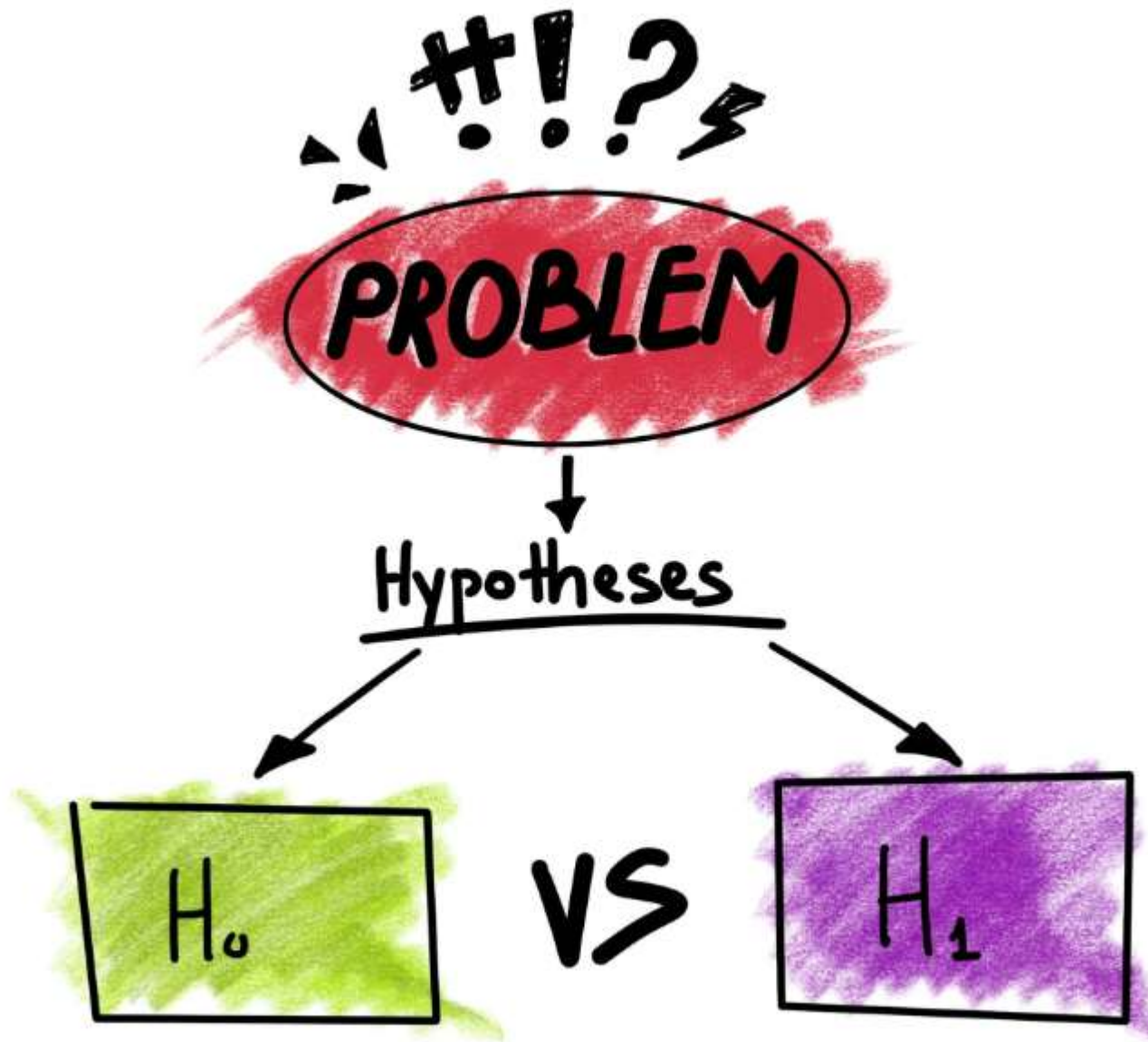
➤ \bar{x} = Mean

➤ x = Random Variable

➤ n = no. of data types

Research Problem Vs Hypothesis

- **Research Problem:** What is the impact of spending a lot of time on mobiles on the attention span of teenagers.
- **Alternative Problem:** Spending time on the mobiles and attention span have a negative correlation.
- **Null Hypothesis:** There does not exist any correlation between the use of mobile by teenagers on their attention span.
- **Research Problem:** What is the impact of providing flexible working hours to the employees on the job satisfaction level.
- **Alternative Hypothesis:** Employees who get the option of flexible working hours have better job satisfaction than the employees who don't get the option of flexible working hours.
- **Null Hypothesis:** There is no association between providing flexible working hours and job satisfaction.



What is a Hypothesis

- In statistics, a hypothesis is a claim or statement about a characteristic of a population.
- A **Null Hypothesis**, denoted H_0 , is a statistical hypothesis that contains a statement of equality such as $\leq, =, \text{or } \geq$.
- The **Alternative Hypothesis**, denoted H_a , is the complement of the null hypothesis. It is a statement that must be true if H_0 is false and it contains a statement of inequality, such as $<, \neq, \text{or } >$.

Rejection of H_0 leads to acceptance of H_a

Age of patients:

The admissions office at Memorial Hospital recently stated that the mean age of its patients was 46 years, with a standard deviation of 20 years. A random sample of 120 ages was obtained from admissions office records in an attempt to disprove the claim. Is a sample mean of 44.2 years significantly smaller than the claimed 46 years at the $\alpha = .10$ level of significance?

Hamburger Fat Content:

A restaurant claims that its hamburgers have no more than 10 grams of fat. You work for a nutritional health agency and are asked to test this claim. You find that a random sample of nine hamburgers has a mean fat content of 13.5 grams and a standard deviation of 5.8 grams. A $\alpha = 0.10$, do you have evidence to reject the restaurant's claim?

Statistical Hypothesis

- *One sample mean test (large sample size)*

Example:

It is claimed that the average weight of a bag of biscuit is 250 grams with a standard deviation of 20.5 grams. Would you agree to this claim if the random sample of 50 bags of biscuits showed an average weight of 240 grams, using a 0.05 level of significance?

- *Two sample mean test (large sample size)*

Example:

Fifty senior students in Statistics got an average of 85 with a standard deviation of 10.2, while a group of 60 senior students have an average of 80 with a standard deviation of 8.9. Can the difference in the mean grade be attributed to chance, using a 0.05 level of significance?

- *One sample mean test small sample size)*
- *Example:*

An expert typist can type 65 words per minute. A random sample of 16 applicants took the typing test and an average speed of 62 words per minute with a standard deviation of 8 words per minute was obtained. Can we say that the applicant's performance is below the standard at 0.05 level of significance

- *Two sample mean test small sample size)*

Example:

A course in Statistics is taught to 12 students by the conventional classroom procedure. A second group of 10 students was given the same course by means of programmed materials. At the end of the term, the same examination was given to each group. The 12 students meeting in the classroom made an average of 85 with a standard deviation of 4, while the 10 students using programmed materials made an average of 81 with a standard deviation of 5. Test the hypothesis that the two methods of learning are equal using 0.10 level of significance. Assume the populations to be approximately normal with equal variances

- *Paired observation*

Example:

The following are the average weekly losses of man-hour due to accidents in 10 industrial plants before and after a certain safety program was put into operation:

45 and 36	33 and 35
73 and 60	57 and 51
46 and 44	83 and 77
124 and 119	34 and 29
26 and 24	17 and 11

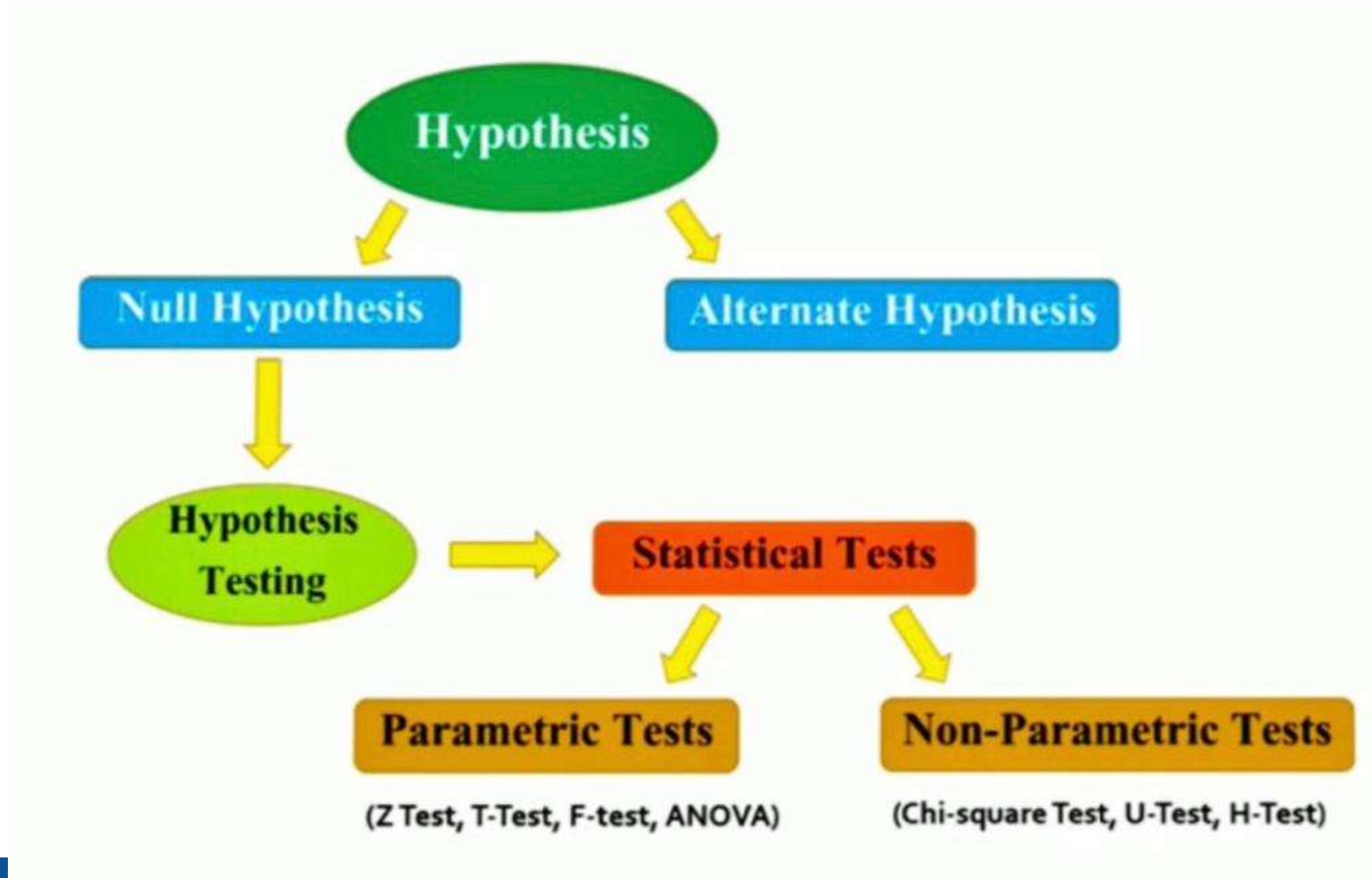
Use 0.05 level of significance to test whether the safety program is effective

Types of Error

- *Type I Error* – Rejection of the null hypothesis when in fact it is true.
- *Type II Error* – Acceptance of the null hypothesis when in fact it is false.

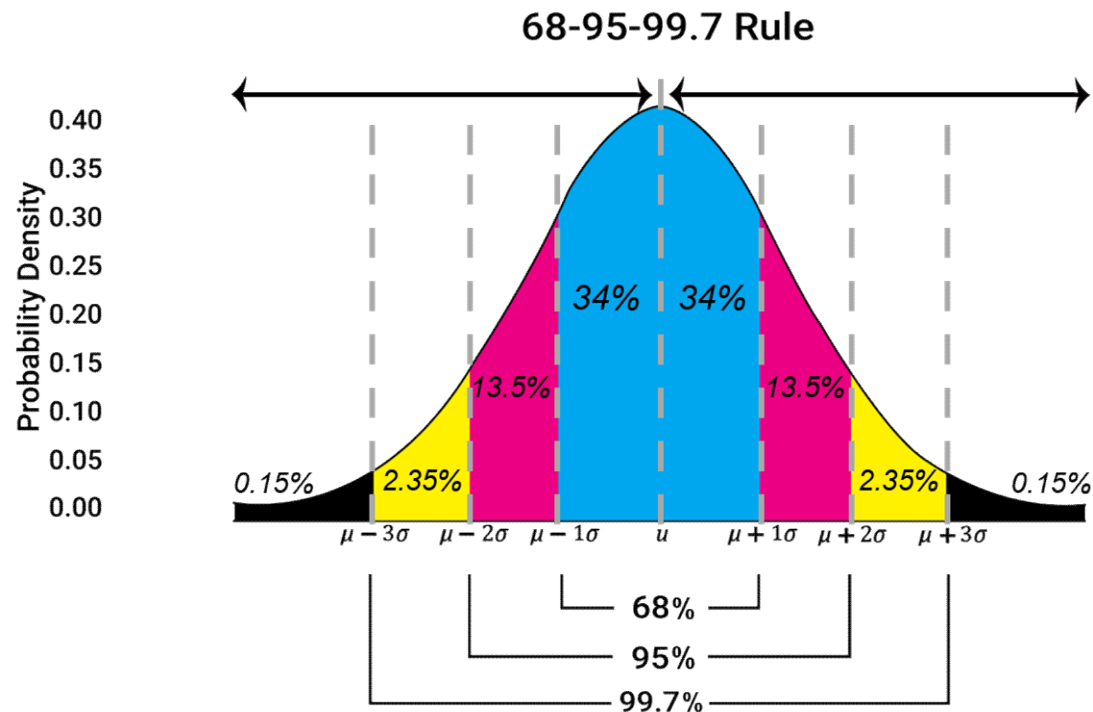
	H_0 is true	H_0 is false
Accept H_0	Correct Decision	Type II Error
Reject H_0	Type I Error	Correct Decision

INFERENCE STATISTICS - Hypothesis Tests


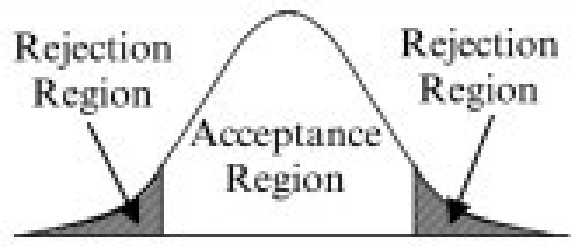



Introduction

- Statistical Hypothesis
- H_0 & H_1
- Possible outcomes
 - No Error
 - Type-I Error (Reject H_0 when it is TRUE)
 - Type-II Error (Do not reject H_0 when it is FALSE)
- Level of Significance
 - $\alpha = P(\text{Type I Error})$
 - $\beta = P(\text{Type II Error})$
- Test Statistics
 - Z-Statistics, t-statistics, F-statistics, Chi-Square test
- Standard Normal Distribution
- Critical Region
 - Region of Rejection, Region of Acceptance
- One-tailed & Two-tailed test



*The bell curve shows you vital statistics of a Normal Distribution,
it is the representation of empirical rule.*

One-Tailed Test (Left Tail)	Two-Tailed Test	One-Tailed Test (Right Tail)
$H_0 : \mu_X = \mu_0$ $H_1 : \mu_X < \mu_0$	$H_0 : \mu_X = \mu_0$ $H_1 : \mu_X \neq \mu_0$	$H_0 : \mu_X = \mu_0$ $H_1 : \mu_X > \mu_0$
		

Z Test Statistic Formula

Test Statistic when σ known

$$z = \frac{\bar{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}} \right)} \%$$

\bar{x} : sample mean
 μ : population mean
 σ : population standard deviation
 n : sample size

T Test Statistic Formula

Test Statistic when σ unknown

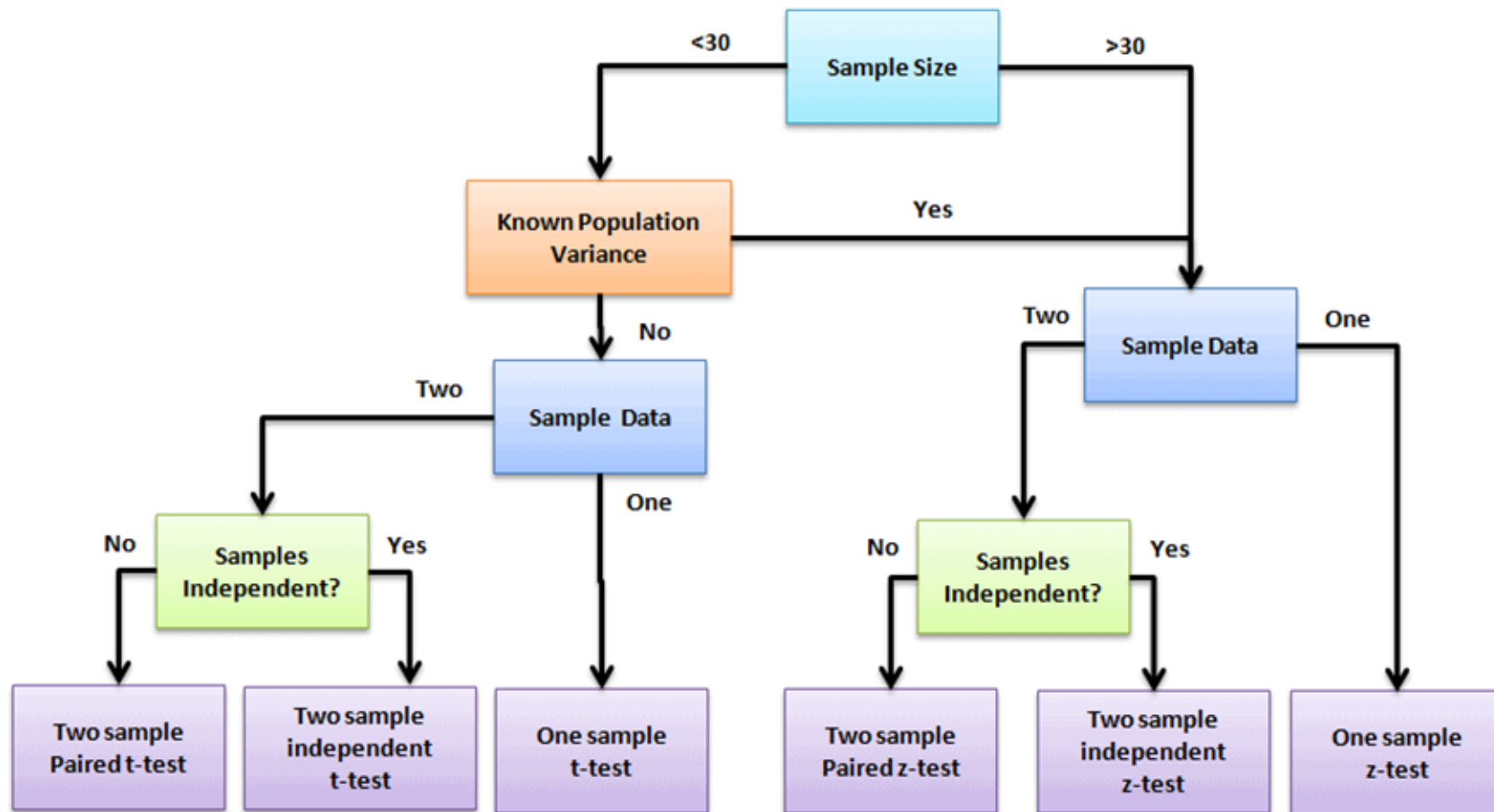
$$t = \frac{\bar{x} - \mu}{\left(\frac{s}{\sqrt{n}} \right)}, \quad df = n - 1$$

\bar{x} : sample mean
 μ : population mean
 s : sample standard deviation
 n : sample size

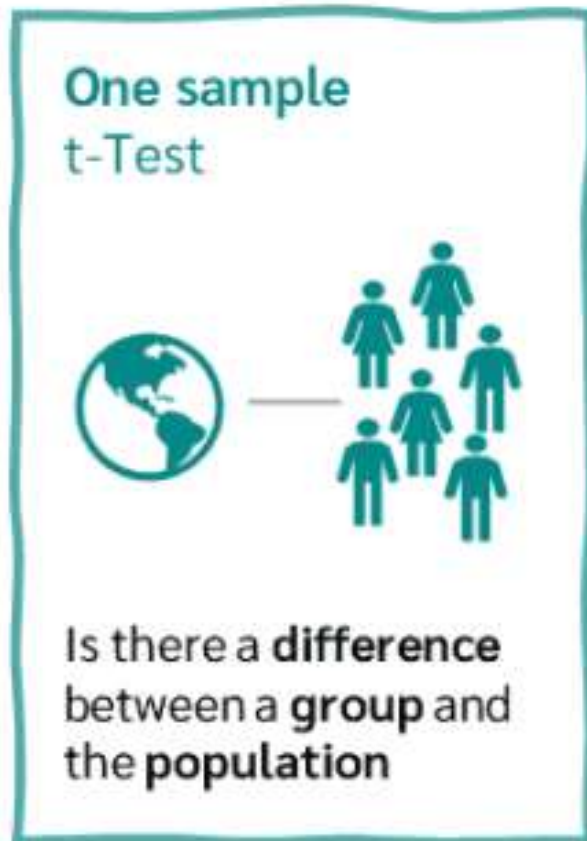
Steps to Hypothesis Test

- State the Hypothesis H_0 & H_1
- Determine the appropriate test statistics
- Set level of significance & critical region
- Compute test statistics using sample data provided
- Make a decision

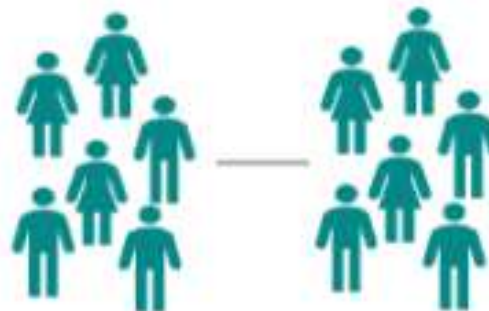
How to identify the Test Statistics



Types of T-Test



Independent samples t-Test



Is there a **difference** between **two groups**

Paired samples t-Test



Is there a **difference** in a **group** between **two points in time**

Student	Score
1	28
2	29
3	35
4	37
5	32
6	26
7	37
8	39
9	22
10	29
11	36
12	38

Significance level

$$\alpha = 0.05$$

Number of sample values

$$n = 12$$

Mean value

$$\bar{x} = 32.33$$

Standard deviation

$$s = 5.47$$

Standard error of the mean

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{5.47}{\sqrt{12}} = 1.58$$

t-value

$$t = \frac{\bar{x} - \mu}{s_{\bar{x}}} = \frac{32.33 - 28}{1.58} = 2.75$$

Degrees of freedom

$$df = n - 1 = 11$$

Example -One Sample T Test

- imagine a company wants to test the claim that their batteries last more than 40 hours. Using a simple random sample of 15 batteries yielded a mean of 44.9 hours, with a standard deviation of 8.9 hours. Test this claim using a significance level of 0.05.

$$H_0: \mu = 40 \quad \hat{x} = 44.9, \quad \mu = 40 \quad s = 8.9, \quad n = 15, \quad df = n - 1 \rightarrow df = 15 - 1 = 14$$

$$H_a: \mu > 40$$

$$\text{test statistic: } t = \frac{44.9 - 40}{\left(\frac{8.9}{\sqrt{15}} \right)} = 2.13$$

- Calculated t 2.13 > table t 1.761, therefore Reject H0. So batteries last more than 40 hrs.

- Suppose a grocery store sells “16 ounce” boxes of Captain Crisp cereal. A random sample of 9 boxes was taken and weighed. The weight in ounces is stored in the data frame `capt-crisp`.

15.5, 16.2, 16.1, 15.8, 15.6, 16.0, 15.8, 15.9, 16.2

- The company that makes Captain Crisp cereal claims that the average weight of a box is at least 16 ounces. We will assume the weight of cereal in a box is normally distributed and use a 0.05 level of significance to test the company's claim.

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

Sixteen oil tins are taken at random from an automatic filling machine. The mean weight of the tins is 14.2 kg, with a standard deviation of 0.40 kg. Can we conclude that the filling machine is wasting oil by filling more than the intended weight of 14 kg, at a significance level of 5%? (Assuming normality).

A random sample of size 16 has 53 as mean and the sum of squares of the deviations taken from the mean is 150. Can the sample be regarded as taken from a normal population with mean 56? (Significance level is 5%)

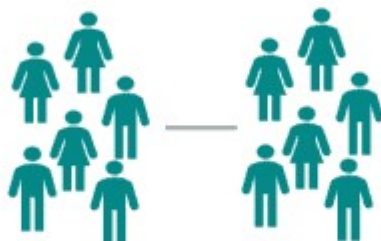
Two(Independent) Samples t-Test

One sample
t-Test



Is there a **difference** between a **group** and the **population**

Independent
samples t-Test

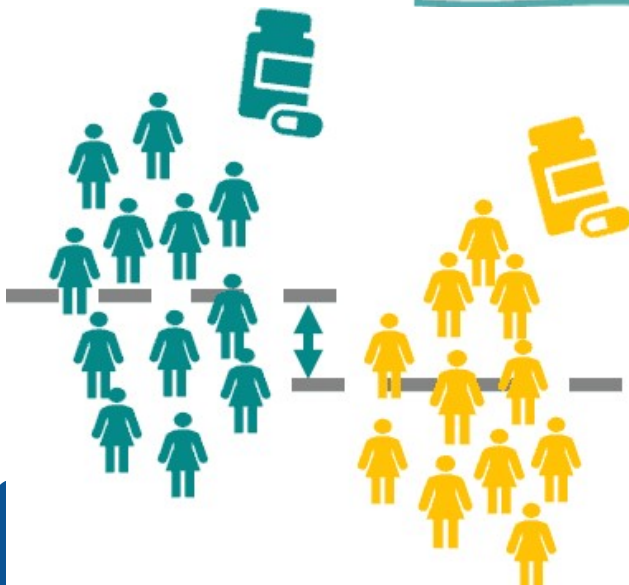


Is there a **difference** between **two groups**

Paired samples
t-Test



Is there a **difference** in a **group** between **two points in time**



- **Null hypothesis:** The means in the two groups are equal (so there is no difference between the two groups).
- **Alternative hypothesis:** The mean values in the two groups are not equal (i.e. there is a difference between the two groups).

Test Statistics

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{\Delta}}}$$

where

$$s_{\bar{\Delta}} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

t is called the t-value,
 \bar{x}_1 and \bar{x}_2 are the means of the two groups which are being compared,
 s_1 and s_2 are the standard deviations of the first and second sets of values
 n_1 and n_2 are the numbers of observations of the first and second groups, respectively.

Example-1

- Is there a significant difference in test scores between 25 students who received in-person instruction and 25 students who received online instruction? The mean test score for the in-person group is 80 (SD = 5) and for the online group is 75 (SD = 7).

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\Delta}}$$

where

$$s_{\Delta} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

- Df = $n_1 + n_2 - 2$ $25 + 25 - 2 = 48$
- a significance level of 0.05 and df = 48.
The critical t-value is 2.01
- $t = (80 - 75) / (\text{sqrt}((5^2/25) + (7^2/25))) = 2.02$

- Since the calculated t-value of 2.02 is greater than the critical t-value of 2.01, we can conclude that there is a significant difference between the test scores of students who receive in-person instruction versus those who receive online instruction

Two(Independent) Samples t -Test

- Twenty participants were given a list of 20 words to process. The 20 participants were randomly assigned to one of two treatment conditions. Half were instructed to count the number of vowels in each word (shallow processing). Half were instructed to judge whether the object described by each word would be useful if one were stranded on a desert island (deep processing). After a brief distractor task, all subjects were given a surprise free recall task. The number of words correctly recalled was recorded for each subject. Here are the data:

Shallow Processing:	13	12	11	9	11	13	14	14	14	15
Deep Processing:	12	15	14	14	13	12	15	14	16	17

- Did the instructions given to the participants significantly affect their level of recall ($\alpha = .05$)?

Example-2

- A researcher wants to know if there is a significant difference in the weight of newborn babies between two hospitals in a city. The researcher randomly selects 20 newborns from Hospital A and 20 newborns from Hospital B and records their weights in pounds. The mean weight for the Hospital A group is 7.5, with a standard deviation of 0.8. The mean weight for the Hospital B group is 7.1, with a standard deviation of 1.2. Is there a significant difference between the two hospitals?
- $t = (7.5 - 7.1) / (\sqrt{(0.8^2/20) + (1.2^2/20)}) = t = 1.77$
- Assuming a significance level of 0.05 and $df = 38$, the critical t-value is 2.024.
- Since the calculated t-value of 1.77 is less than the critical t-value of 2.024, we can conclude that there is not a significant difference in the weight of newborn babies between the two hospitals in the city.

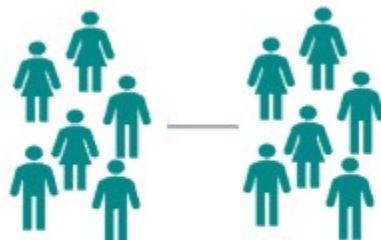
Two Sample(Dependent) Paired t-Test

One sample
t-Test



Is there a **difference** between a **group** and the **population**

Independent
samples t-Test

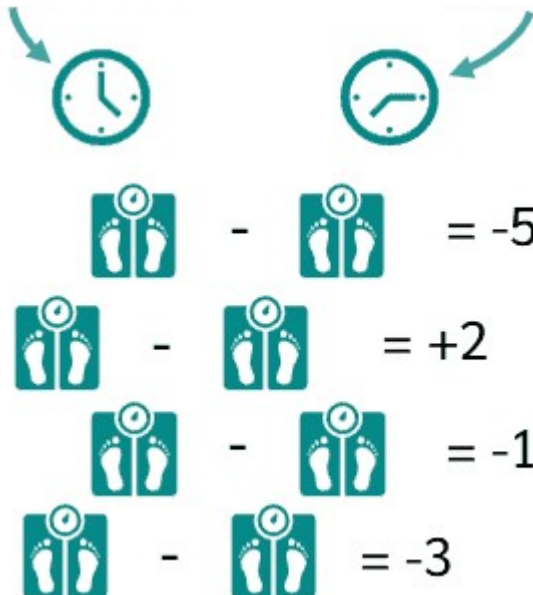


Is there a **difference** between **two groups**

Paired samples
t-Test



Is there a **difference** in a **group** between **two points in time**



- **Null hypothesis:** The mean of the differences between the pairs is zero.
- **Alternative hypothesis:** The mean of the differences between the pairs is non-zero.

- The water diet requires you to drink 2 cups of water every half hour from when you get up until you go to bed but eat anything you want. Four adult volunteers agreed to test this diet. They are weighed prior to beginning the diet and 6 weeks after. Their weights in pounds are

Person	1	2	3	4	mean	S.d.
Weight before	180	125	240	150	173.75	49.56
Weight after	170	130	215	152	166.75	36.09
Difference	10	-5	25	-2	7	13.64

- $H_0: \text{Diff} = 0$ (no difference -- there is no difference at all in the mean of the weight difference)
 - $H_a: \text{Diff} \neq 0$ (difference -- diet made difference in the means of the weight differences)
- From the data, we know $\overline{\text{Diff}} = 7$ and $s_{\text{Diff}} = 13.64$. Then we get

$$t = \frac{\overline{\text{Diff}} - \mu_0}{\frac{s_{\text{Diff}}}{\sqrt{n}}} = \frac{7 - 0}{\frac{13.64}{\sqrt{4}}} = \frac{7}{6.82} = 1.026.$$

- we fail to discard the null hypothesis at the 0.05 significance level. We do not have enough confirmation to deduce that the water diet has an impact on the weight

References

- <https://datatab.net/tutorial/t-test>
- <https://collegedunia.com/exams/t-test-formula-mathematics-articleid-4856>

Testing the Hypothesis

- Basic Approaches
 - P value approach
 - Critical value approach

P – Value Approach

A p-value is the lowest level (of significance) at which the observed value of the test statistic is significant

- **PROCEDURES FOR HYPOTHESIS TESTING**

- State the null and alternative hypothesis.
- Choose the level of significance α
- Determine the test to be used, t or z test, one or two-tailed test.
- Determine the critical region.
- Compute the value of the test statistic from the sample data.
- Make decision whether to accept or reject the null hypothesis

- **NOTE: REJECT** the null hypothesis H_0 if the test statistic has a value in the critical region (or if the computed p-value is less than or equal to the desired level of significance level, α otherwise, **ACCEPT** H_0

One Sample t-Test

- Suppose a grocery store sells “16 ounce” boxes of Captain Crisp cereal. A random sample of 9 boxes was taken and weighed. The weight in ounces is stored in the data frame `capt-crisp`.

15.5, 16.2, 16.1, 15.8, 15.6, 16.0, 15.8, 15.9, 16.2

- The company that makes Captain Crisp cereal claims that the average weight of a box is at least 16 ounces. We will assume the weight of cereal in a box is normally distributed and use a 0.05 level of significance to test the company's claim.

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

```
> capt_crisp = data.frame(weight = c(15.5, 16.2, 16.1,
+                                15.8, 15.6, 16.0,
+                                15.8, 15.9, 16.2))
> x_bar = mean(capt_crisp$weight)
> s = sd(capt_crisp$weight)
> mu_0 = 16
> n = 9
> t = (x_bar - mu_0) / (s / sqrt(n))
> t
[1] -1.2
> #P(t8 < -1.2)
> pt(t, df = n - 1)
[1] 0.1322336
> t.test(x = capt_crisp$weight, mu = 16, alternative = c("less"), conf.level = 0.95)
```

One Sample t-test

```
data: capt_crisp$weight
t = -1.2, df = 8, p-value = 0.1322
alternative hypothesis: true mean is less than 16
95 percent confidence interval:
 -Inf 16.05496
sample estimates:
mean of x
 15.9

> # two-sided interval for the mean weight of boxes of Captain Crisp cereal
> capt_test_results = t.test(capt_crisp$weight, mu = 16,
+                           alternative = c("two.sided"),
+                           conf.level = 0.95)
> names(capt_test_results)
[1] "statistic" "parameter" "p.value" "conf.int" "estimate" "null.val
ue" "stderr" "alternative" "method" "data.name"
> capt_test_results$method
[1] "One Sample t-test"
```

Two sample(Independent) Test

- Assume that the distributions of X and Y are $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$, respectively. Given the $n = 6$ observations of X ,
- we will test $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 > \mu_2$.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\Delta}}$$

where

$$s_{\Delta} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

```
> x = c(70, 82, 78, 74, 94, 82)
> n = length(x)
> y = c(64, 72, 60, 76, 72, 80, 84, 68)
> m = length(y)
> x_bar = mean(x)
> s_x = sd(x)
> y_bar = mean(y)
> s_y = sd(y)
> s_p1=sqrt((s_x^2/n)+(s_y^2/m))
> s_p1
[1] 4.412105
> t=(x_bar-y_bar)/s_p1
> t
[1] 1.813194
> 1 - pt(t, df = n + m - 2)
[1] 0.0474371
> #Using t-Test function
> t.test(x, y, alternative = c("greater"), var.equal = TRUE)
```

Two Sample t-test

```
data: x and y
t = 1.8234, df = 12, p-value = 0.04662
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.1802451      Inf
sample estimates:
mean of x mean of y
      80      72
```

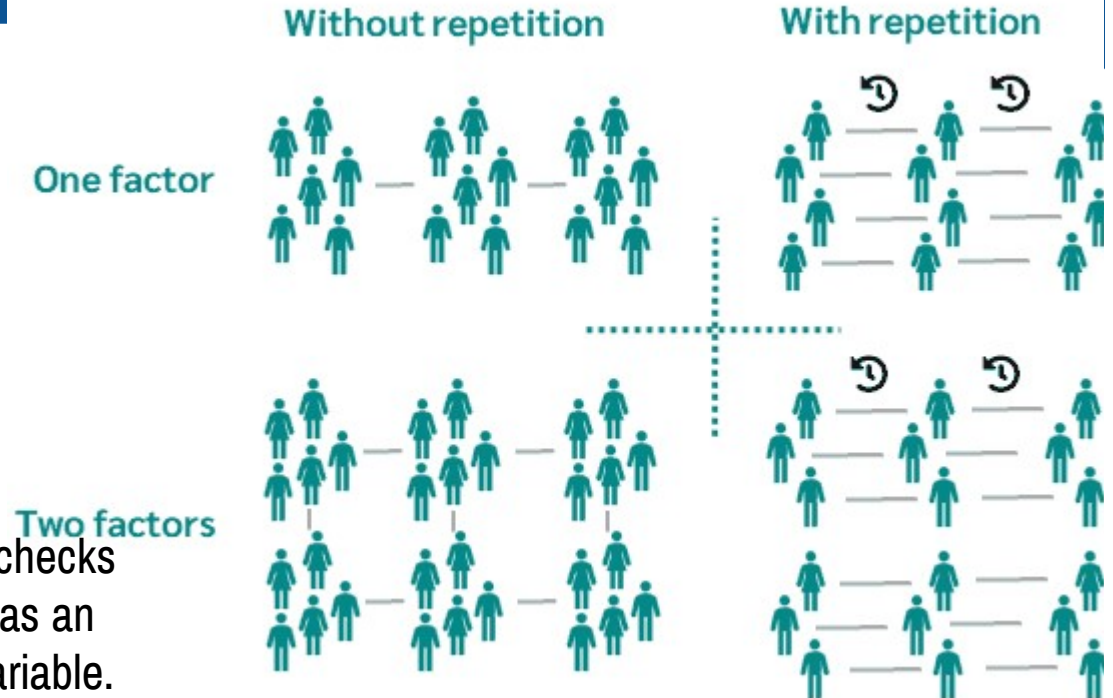
Two Sample Paired Test

Analysis of Variance(ANOVA)

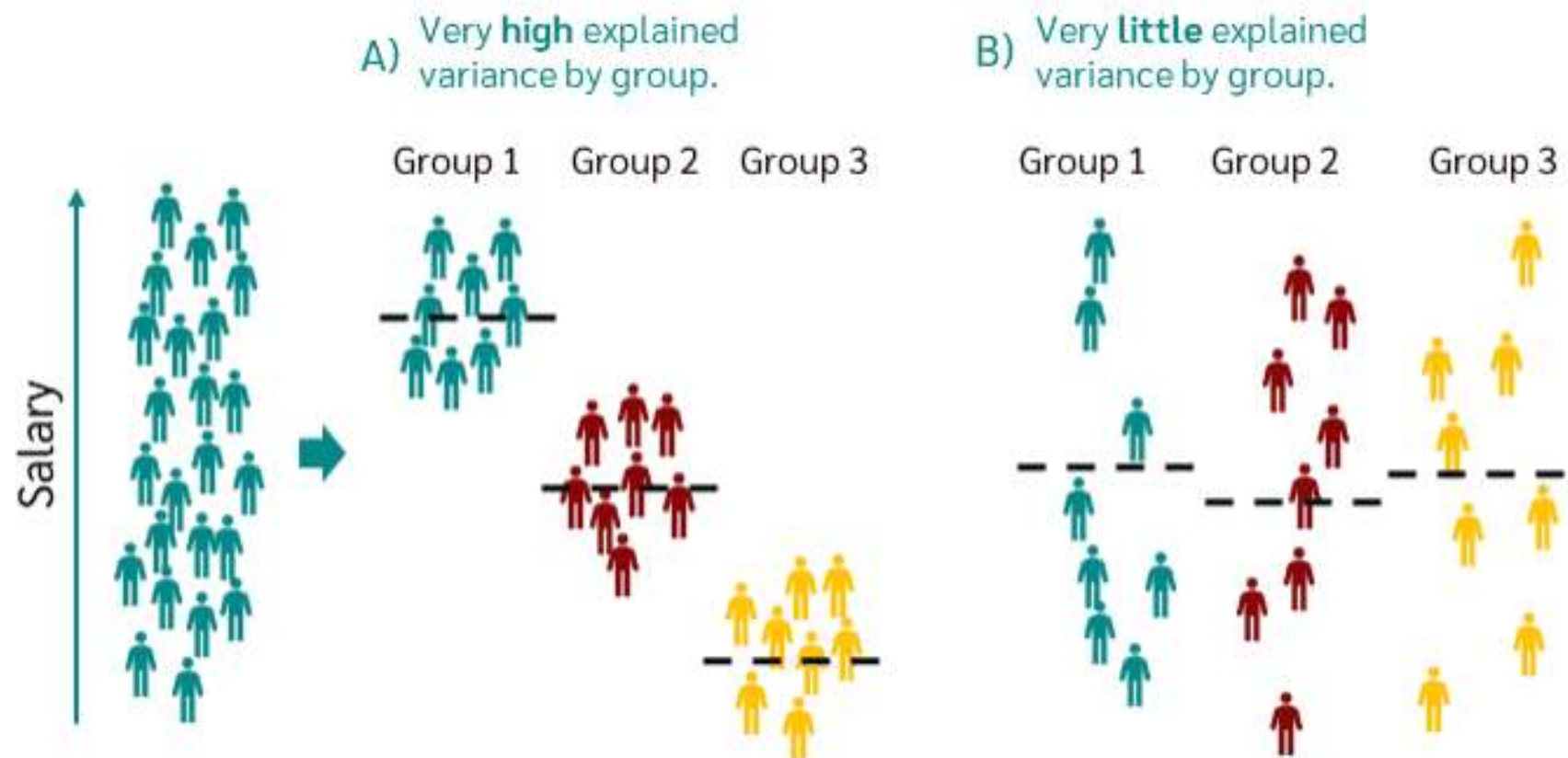
- ANOVA tests whether statistically significant differences exist between more than two samples.
- Means and variances of the respective groups are compared with each other.
- Types of ANOVA
 - One-factor (or one-way) ANOVA
 - Two-factors (or two-way) ANOVA
 - One-factor ANOVA with repeated measurements
 - Two-factors ANOVA with repeated measurements
- Why not calculate multiple t-tests?
 - This probability of error is usually set at 5%, so that, from a purely statistical point of view, every 20th test gives a wrong result
- Types of Variations
 - Between samples
 - Within sample

Types of ANOVA

- Independent Variable
 - Water, Fertilizer, Sunlight
- Dependent variable
 - Growth
- ANOVA
 - one-way analysis of variance only checks whether an independent variable has an influence on a metric dependent variable.
 - Does a person's place of residence (independent variable) influence his or her salary?
 - Two-way analysis has 2 independent Variables
 - Does a person's place of residence (1st independent variable) and gender (2nd independent variable) affect his or her salary?

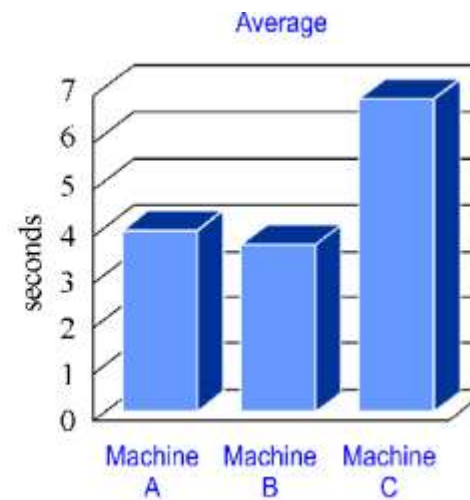
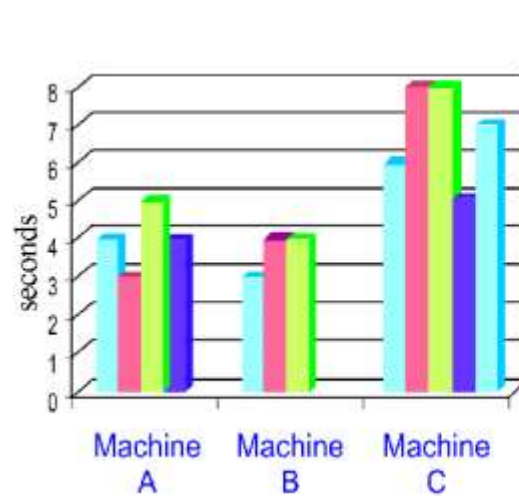


Variation within Sample Vs Variation between Sample



	Variance within the groups	Variance between group means
Case A)	Small	Large
Case B)	Large	Small

	Time Taken in Seconds					Average
Machine A	4	3	5	4		4
Machine B	3	4	4			3.66
Machine C	6	8	8	5	7	6.8



One-Way & Two-Way ANOVA

	Dependent variable	Independent variable
Level of measurement	An metric-scaled variable	A nominally scaled variable with more than two levels
Example	Weekly coffee consumption	Subject (math, psychology, economics)

	Dependent variable	Independent variable
Level of measurement	One metric-scaled variable	Two nominally scaled variables
Example	Weekly coffee consumption	Subject (math, psychology, economics) and semester (winter, summer)

Assumptions of ANOVA

- Normality
 - The population from which the various samples are selected are normally distributed
- Homogeneity
 - The populations from which the samples are drawn have the same variance
- Independence
 - The samples are independently drawn
- Additivity
 - The effect of various components are additive

	Dependent variable	Independent variable
Level of measurement	An metric-scaled variable	A nominally scaled variable with more than two levels
Example	Weekly coffee consumption	Subject (math, psychology, economics)

One-Way ANOVA examples

- Company has to study whether the six varieties of products marketed by the company have equal demand in the market or not
- Different teaching methods have equal efficiency or not
- Efficiency of 3 fertilisers are significantly different or not
- Five different drugs produced for a particular disease are equally efficient or not

One-Way ANOVA Procedure

	Time Taken in Seconds					Total	
Machine A	4	3	5	4		16	(T ₁)
Machine B	3	4	4			11	(T ₂)
Machine C	6	8	8	5	7	34	(T ₃)
	Total					61	(G)

Step I

The null hypothesis is

H_0 : The means are equal

Or $H_0: \mu_1 = \mu_2 = \mu_3 = \dots = \mu_k$

against the alternative hypothesis

H_1 : The means are not equal

number of plots = $n_1 + n_2 + n_3 + \dots + n_k = N$. (In the above example: $k=3$. $n_1 = 4$, $n_2 = 3$ and $n_3 = 5$. Therefore $N = 4 + 3 + 5 = 12$.)

Step II

- 1) Compute the Correction Factor, $CF = \frac{G^2}{N}$

Where G is the grand total.

- 2) Compute Total Sum of Squares, $TSS = \text{sum of squares of all observations} - CF$.
- 3) Compute Between Sum of Squares (SSB) or Treatment Sum of Squares (SST)

$$SSB = \sum \frac{T_i^2}{n_i} - CF \text{ where } T_i \text{ the sum of observations in the } i^{\text{th}} \text{ sample,}$$

n_i is the number of observations in the i^{th} sample

- 4) Compute Within Sum of Squares (SSW) or Error Sum of Squares (SSE)

$$SSW = TSS - SSB$$

Step III

Compute Mean Sum of Squares

$$\text{Mean sum of squares} = \frac{\text{Sum of squares}}{\text{Degrees of freedom}}$$

Here degrees of freedom for TSS = $N - 1$
 degrees of freedom for SSB = $k - 1$
 degrees of freedom for SSW = $N - k$

Therefore, Mean Between Sum of Squares is, $MSB = \frac{SSB}{k - 1}$

Mean Within Sum of Squares is, $MSW = \frac{SSW}{N - k}$

Step IV

Compute F ratio

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

Step V

Draw the ANOVA Table.

The format of ANOVA Table is shown below.

Step VI

If computed F is greater than F_{α} (i.e. $F > F_{\alpha}$) for a given level of significance α , we reject the null hypothesis, otherwise we accept H_0 .

ANOVA Table

Source	df	Sum of Squares (SS)	Mean Sum of Squares (MSS)	F	F_{α}
Between Samples (Treatment)	$k - 1$	SSB	MSB	$F = \frac{MSB}{MSW}$	$F(k-1, N-k)$
Within Samples (Error)	$N - k$	SSW	MSW		
Total	$N - 1$	TSS			

Example-1

- The time taken in seconds by three different packing machines is given below. Test whether the machines are equally efficient or not at 5% level of significance.

Time Taken (in Seconds)					
Machine A	4	3	5	4	
Machine B	3	4	4		
Machine C	6	8	8	5	7

	Time taken in seconds					Total	
Machine A	4	3	5	4		16	(T ₁)
Machine B	3	4	4			11	(T ₂)
Machine C	6	8	8	5	7	34	(T ₃)
	Total					61	(G)

Two-Way ANOVA

- two-way ANOVA a dependent variable is compared over three or more groups, controlling for another variable.
- The grouping is taken as one factor and the control is taken as another factor. The grouping factor is usually known as Treatment. The control factor is usually called Block.
- The accuracy of the test in two -way ANOVA is considerably higher than that of the one-way ANOVA, as the additional factor, block is used to reduce the error variance.

Groups or Treatments	Blocks					
		1	2	3	...	m
	1	x_{11}	x_{12}	x_{13}	...	x_{1m}
	2	x_{21}	x_{22}	x_{23}	...	x_{2m}
	3	x_{31}	x_{32}	x_{33}	...	x_{3m}
	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
	k	x_{k1}	x_{k2}	x_{k3}	...	x_{km}
	$x_{.j}$	$x_{.1}$	$x_{.2}$	$x_{.3}$...	$x_{.m}$
						G

	Water temperature		
Detergent	Cold	Warm	Hot
Detergent x	4	7	10
	5	8	11
	5	9	12
	6	12	19
	5	3	15
Detergent y	4	12	10
	4	12	12
	6	13	13
	6	15	13
	5	13	12

		Oven Temperature		
		325	350	400
Type of Sugar	white sugar	10.75	8.75	4.00
		9.50	8.25	5.50
		10.00	9.00	4.75
		10.00	8.00	4.00
		9.25	8.25	5.00
	white & brown sugar	12.00	10.25	7.00
		10.00	9.00	7.25
		10.50	8.50	6.50
		11.25	10.50	5.00
		11.00	9.75	8.00

Two-Way ANOVA Procedure

Step 1 : In two-way ANOVA we have two pairs of hypotheses, one for treatments and one for the blocks.

Framing Hypotheses

Null Hypotheses

H_{01} : There is no significant difference among the population means of different groups (Treatments)

H_{02} : There is no significant difference among the population means of different Blocks

Alternative Hypotheses

H_{11} : Atleast one pair of treatment means differs significantly

H_{12} : Atleast one pair of block means differs significantly

Step 2 : **Data** is presented in a rectangular table form as described in the previous section.

Step 3 : **Level of significance** α .

Step 4 : Test Statistic

$$F_{0t}(\text{treatments}) = MST / MSE$$

$$F_{0b}(\text{block}) = MSB / MSE$$

To find the test statistic we have to find the following intermediate values.

i) Correction Factor: $C.F = \frac{G^2}{n}$ where $G = \sum_{j=1}^m \sum_{i=1}^k x_{ij}$

ii) Total Sum of Squares: $TSS = \sum_{i=1}^k \sum_{j=1}^m x_{ij}^2 - C.F$ vi) Degrees of freedom

iii) Sum of Squares between Treatments: $SST = \sum_{i=1}^k \frac{x_{i.}^2}{m} - C.F$

iv) Sum of squares between blocks: $SSB = \sum_{j=1}^m \frac{x_{.j}^2}{k} - C.F$

Degrees of freedom (d.f.)	d.f.
Total Sum of Squares	$n-1$
Treatment Sum of Squares	$k-1$
Block Sum of Squares	$m-1$
Error of Sum Squares	$(m-1)(k-1)$

v) Sum of Squares due to Error: $SSE = TSS - SST - SSB$

vii) Mean Sum of Squares

Mean sum of Squares due to Treatments: $MST = \frac{SST}{k-1}$

Mean sum of Squares due to Blocks: $MSB = \frac{SSB}{m-1}$

Mean sum of Squares due to Error: $MSE = \frac{SSE}{(k-1)(m-1)}$

Step 5 : Calculation of the Test Statistic

ANOVA Table (two-way)

Source of variation	Sum of squares	Degrees of freedom	Mean sum of squares	F-ratio
Treatments	SST	$k-1$	MST	$F_{0t} = \frac{MST}{MSE}$
Blocks	SSB	$m-1$	MSB	$F_{0b} = \frac{MSB}{MSE}$
Error	SSE	$(k-1)(m-1)$	MSE	
Total	TSS	$n-1$		

Step 6 : Critical values

Critical value for treatments = $f_{(k-1, (m-1)(k-1)), \alpha}$

Critical value for blocks = $f_{(m-1, (m-1)(k-1)), \alpha}$

Step 7 : Decision

For Treatments: If the calculated F_{0t} value is greater than the corresponding critical value, then we reject the null hypothesis and conclude that there is significant difference among the treatment means, in atleast one pair.

For Blocks: If the calculated F_{0b} value is greater than the corresponding critical value, then we reject the null hypothesis and conclude that there is significant difference among the block means, in at least one pair.

Example -1

A reputed marketing agency in India has three different training programs for its salesmen. The three programs are Method – A, B, C. To assess the success of the programs, 4 salesmen from each of the programs were sent to the field. Their performances in terms of sales are given in the following table. Test whether there is significant difference among methods and among salesmen.

Salesmen	Methods		
	A	B	C
1	4	6	2
2	6	10	6
3	5	7	4
4	7	5	4

Step 1 : Hypotheses

Null Hypotheses: $H_{01} : \mu_{M1} = \mu_{M2} = \mu_{M3}$ (for treatments)

That is, there is no significant difference among the three programs in their mean sales.

$H_{02} : \mu_{S1} = \mu_{S2} = \mu_{S3} = \mu_{S4}$ (for blocks)

Alternative Hypotheses:

H_{11} : At least one average is different from the other, among the three programs.

H_{12} : At least one average is different from the other, among the four salesmen.

Step 2 : Data

Salesmen	Methods		
	A	B	C
1	4	6	2
2	6	10	6
3	5	7	4
4	7	5	4

Step 3 : Level of significance $\alpha = 5\%$

Step 4 : Test Statistic

$$F_{0t}(\text{treatment}) = \frac{MST}{MSE}$$

$$F_{0b}(\text{block}) = \frac{MSB}{MSE}$$

Step-5 : Calculation of the Test Statistic

	Methods			Total x_i	x_i^2
	A	B	C		
1	4	6	2	12	144
2	6	10	6	22	484
3	5	7	4	16	256
4	7	5	4	16	256
x_i	22	28	16	66	1140
x_i^2	484	784	256	1524	

Squares

16	36	4
36	100	36
25	49	16
49	25	16
		$\sum_i \sum_{ii} x_{ii}^2 = 408$

Correction Factor: $CF = \frac{G^2}{n} = \frac{(66)^2}{12} = \frac{4356}{12} = 363$

Total Sum of Squares: $TSS = \sum \sum x_{ij}^2 - C.F$
 $= 408 - 363 = 45$

Sum of Squares due to Treatments: $SST = \frac{\sum_{j=1}^k x_{.j}^2}{k} - C.F$
 $= \frac{1140}{3} - 363$
 $= 380 - 363 = 17$

Sum of Squares due to Blocks: $SSB = \frac{\sum_{i=1}^k x_{.j}^2}{k} - C.F$
 $= \frac{1524}{4} - 363$
 $= 381 - 363$
 $= 18$

Sum of Squares due to Error: $SSE = TSS - SST - SSB$
 $= 45 - 17 - 18 = 10$

Mean sum of squares: $MST = \frac{SST}{k-1} = \frac{17}{3} = 5.67$
 $MSB = \frac{SSB}{m-1} = \frac{18}{2} = 9$
 $MSE = \frac{SSE}{(k-1)(m-1)} = \frac{10}{6} = 1.67$

ANOVA Table (two-way)

Sources of variation	Sum of squares	Degrees of freedom	Mean sum of squares	F-ratio
Between treatments (Programs)	17	3	5.67	$F_{ot} = \frac{5.67}{1.67} = 3.40$
Between blocks (Salesmen)	18	2	9	$F_{ob} = \frac{9}{1.67} = 5.39$
Error	10	6	1.67	
Total		11		

Step 6 : Critical values

$$f(3, 6), 0.05 = 4.7571 \text{ (for treatments)}$$

$$f(2, 6), 0.05 = 5.1456 \text{ (for blocks)}$$

Step 7 : Decision

(i) Calculated $F_{ot} = 3.40 < f(3, 6), 0.05 = 4.7571$, the null hypothesis is not rejected and we conclude that there is significant difference in the mean sales among the three programs.

(ii) Calculate $F_{ob} = 5.39 > f(2, 6), 0.05 = 5.1456$, the null hypothesis is rejected and conclude that there does not exist significant difference in the mean sales among the four salesmen.

References

- More Problems for Practice
 - <https://www.brainkart.com/article/Two-Way-ANOVA-39242/>
 - <https://atozmath.com/example/CONM/Anova.aspx?q=anova2&q1=E1>
- Two-Way ANOVA using R
 - <https://statsandr.com/blog/two-way-anova-in-r/>
 - <https://www.r-bloggers.com/2022/05/two-way-anova-example-in-r-quick-guide/>
 - <https://statdoe.com/two-way-anova-in-r/>