RESEARCH METHODOLOGY

Unit-03:
Testing of hypotheses

Dr.Roopa Ravish
Department of CSE

RESEARCH METHODOLOGY

Topic: Basic concepts - Procedure for hypothesis testing, flow diagram for hypothesis testing

Testing of Hypotheses I: Introduction

- Hypothesis is usually considered as the Principal Instrument of research
- Function is to suggest new experiments and observation.

- Hypothesis testing is often used strategy for deciding whether sample data offer such support for hypothesis that generalization can be made.

- Ordinarily, when one talks about hypothesis, one simply means a mere assumption or some supposition to be proved or disproved. But for a researcher hypothesis is a formal question that he intends to resolve.

What is hypothesis testing?

- Mere assumption or some supposition to be proved or disproved.
- Defined as a

"*Proposition or a set of proposition set forth as an explanation for the occurrence of some specified group of phenomena either asserted merely as a provisional conjecture to guide some investigation or accepted as highly probable in the light of established facts.*"

Quite often a research hypothesis is a predictive statement, capable of being tested by scientific methods, that relates an independent variable to some dependent variable.

Examples

a. “Students who receive counselling will show a greater increase in creativity than students not receiving counselling”

b. “The automobile A is performing as well as automobile B”.

These are hypotheses capable of being objectively verified and tested. Thus, we may conclude that a hypothesis states what we are looking for and it is a proposition which can be put to a test to determine its validity.

Characteristics of Hypothesis

1) Should be clear and precise.
2) Should be capable of being tested.
    a) A Hypotheses is testable if other deductions can be made from it which, in turn, can be confirmed or disproved by observation.
3) Should state relationship between variables.
4) Should be limited in scope and must be specific.
5) Hypo should be stated in simple terms and easily understandable.
6) Hypo should be consistent with most known facts.
7) Hypo should be amenable to testing within reasonable time.

Basic concepts: Null Hypothesis and Alternate Hypothesis

In context of Statistical Analysis:

Null Hypothesis – If we compare method A and method B and both are equally good (H_0).

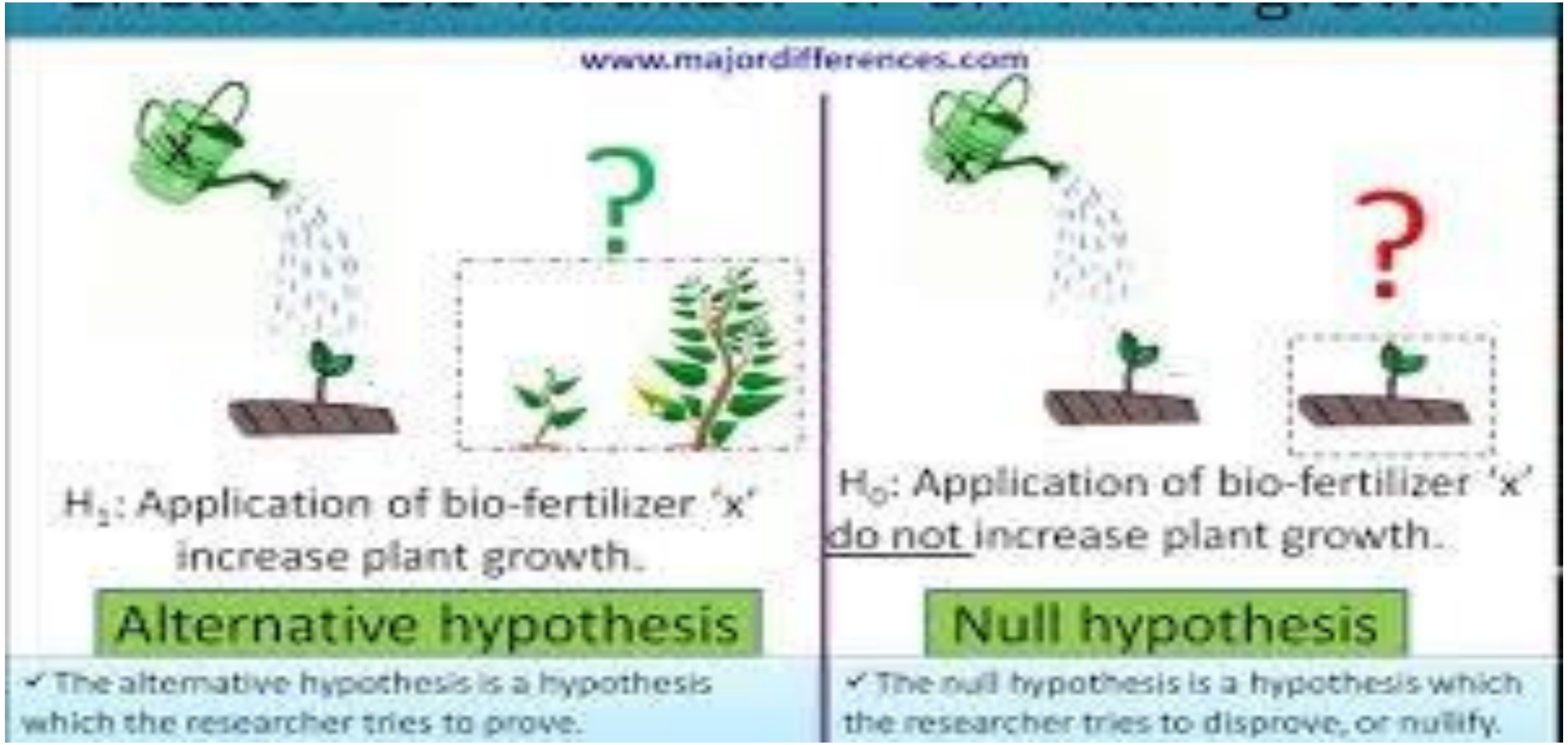
Example *: "No difference between coke and diet coke".*

As against this, we may think that the method A is superior or the method B is inferior, we are then stating what is termed as alternative hypothesis. The null hypothesis is generally symbolized as H0 and the alternative hypothesis as Ha.

Alternate Hypothesis – If method A is superior than B (H_1).

Example *: "There is difference between coke and diet coke".*

Example


www.majordifferences.com
?
?
H_1: Application of bio-fertilizer 'x' increase plant growth.
Alternative hypothesis
✓ The alternative hypothesis is a hypothesis which the researcher tries to prove.
H_0: Application of bio-fertilizer 'x' do not increase plant growth.
Null hypothesis
✓ The null hypothesis is a hypothesis which the researcher tries to disprove, or nullify.

Null Hypothesis

- Suppose we want to test the hypothesis that the population mean (μ) is equal to the hypothesized mean (μ_{H0}) = 100.

- Then we would say that the null hypothesis is that the population mean is equal to the hypothesized mean 100 and symbolically we can express as:

$$H_0: \mu = \mu_{H0} = 100$$

If our sample results do not support this null hypothesis, we should conclude that something else is true.

What we conclude rejecting the null hypothesis is known as alternative hypothesis. Set of alternatives to the null hypothesis is referred to as the alternative hypothesis. If we accept H_0 , then we are rejecting Ha and

If we reject H_0 , then we are accepting Ha.

Possible alternate hypothesis

For $H_0: \mu = \mu_{H0} = 100$, we may consider three possible alternative hypotheses as follows* :

Table 9.1

Alternative hypothesis	*To be read as follows*
$H_a : \mu \neq \mu_{H_0}$	(The alternative hypothesis is that the population mean is not equal to 100 i.e., it may be more or less than 100)
$H_a : \mu > \mu_{H_0}$	(The alternative hypothesis is that the population mean is greater than 100)
$H_a : \mu < \mu_{H_0}$	(The alternative hypothesis is that the population mean is less than 100)

Possible alternate hypothesis

In the choice of null hypothesis, the following considerations are usually kept in view:

(a) Alternative hypothesis is usually the one which one wishes to prove and the null hypothesis is the one which one wishes to disprove.

Thus, a null hypothesis represents the hypothesis we are trying to reject, and alternative hypothesis represents all other possibilities.

(b) If the rejection of a certain hypothesis when it is actually true involves great risk, it is taken as null hypothesis because then the probability of rejecting it when it is true is α (the level of significance) which is chosen very small.

(c) Null hypothesis should always be specific hypothesis i.e., it should not state about or approximately a certain value.

Statistically Significant

- Measurements are done on the two categorical variables on a *sample* of individuals from a population, and they are interested in whether or not there is a relationship between the two variables in the *population.*

- a relationship as strong as the one observed in the sample (or stronger) would be unlikely without a real relationship in the population, then the relationship in the sample is said to be statistically significant.

- The notion that it could have happened just by chance is deemed to be implausible.

Level of Significance

This is a very important concept in the context of hypothesis testing. It is always some percentage (usually 5%) which should be chosen with great care, thought and reason.

In case we take the significance level at 5 per cent, then this implies that H0 will be rejected when the sampling result (i.e., observed evidence) has a less than 0.05 probability of occurring if H0 is true.

In other words, the 5% level of significance means that researcher is willing to take as much as a 5% risk of rejecting the null hypothesis when it (H0) happens to be true.

Thus the significance level is the maximum value of the probability of rejecting H0 when it is true and is usually determined in advance before testing the hypothesis.

Decision rule or test of hypothesis

Given a hypothesis H0 and an alternative hypothesis Ha , we make a rule which is known as decision rule according to which we accept H0 (i.e., reject Ha) or reject H0 (i.e., accept Ha).

Example: If H0 is that a certain lot is good (there are very few defective items in it) against Ha , that the lot is not good (there are too many defective items in it), then we must decide the number of items to be tested and the criterion for accepting or rejecting the hypothesis.

We might test 10 items in the lot and plan our decision saying that if there are none or only 1 defective item among the 10, we will accept H0 otherwise we will reject H0 (or accept Ha). This sort of basis is known as decision rule

Decision rule or test of hypothesis

Given a hypothesis H0 and an alternative hypothesis Ha , we make a rule which is known as decision rule according to which we accept H0 (i.e., reject Ha) or reject H0 (i.e., accept Ha).

Example: If H0 is that a certain lot is good (there are very few defective items in it) against Ha , that the lot is not good (there are too many defective items in it), then we must decide the number of items to be tested and the criterion for accepting or rejecting the hypothesis.

We might test 10 items in the lot and plan our decision saying that if there are none or only 1 defective item among the 10, we will accept H0 otherwise we will reject H0 (or accept Ha). This sort of basis is known as decision rule

Type I and Type II errors

Type I error is denoted by α (alpha) known as α error, also called the level of significance of test; and Type II error is denoted by β (beta) known as β error.

	Decision	
	Accept H_0	Reject H_0
H_0 (true)	Correct decision	Type I error (α error)
H_0 (false)	Type II error (β error)	Correct decision

Type I and Type II errors

The probability of Type I error is usually determined in advance and is understood as the level of significance of testing the hypothesis.

If type I error is fixed at 5 per cent, it means that there are about 5 chances in 100 that we will reject H0 when H0 is true.

We can control Type I error just by fixing it at a lower level. For instance, if we fix it at 1 per cent, we will say that the maximum probability of committing Type I error would only be 0.01.

Type I and Type II errors

But with a fixed sample size, n, when we try to reduce Type I error, the probability of committing Type II error increases. Both types of errors cannot be reduced simultaneously.

There is a trade-off between two types of errors.

To deal with this trade-off in business situations, decision-makers decide the appropriate level of Type I error by examining the costs or penalties attached to both types of errors.

Hence, in the testing of hypothesis, one must make all possible effort to strike an adequate balance between Type I and Type II errors.

One tailed and two tailed test

We test 3 types of Hypotheses given by:

1) $H_0:\mu = \mu_{H0}$ Aganist $H_a:\mu \neq \mu_{H0}$

2) $H_0:\mu = \mu_{H0}$ Aganist $H_a:\mu > \mu_{H0}$

or

$H_0:\mu <= \mu_{H0}$ Aganist $H_a:\mu > \mu_{H0}$

3) $H_0:\mu = \mu_{H0}$ Aganist $H_a:\mu < \mu_{H0}$

or

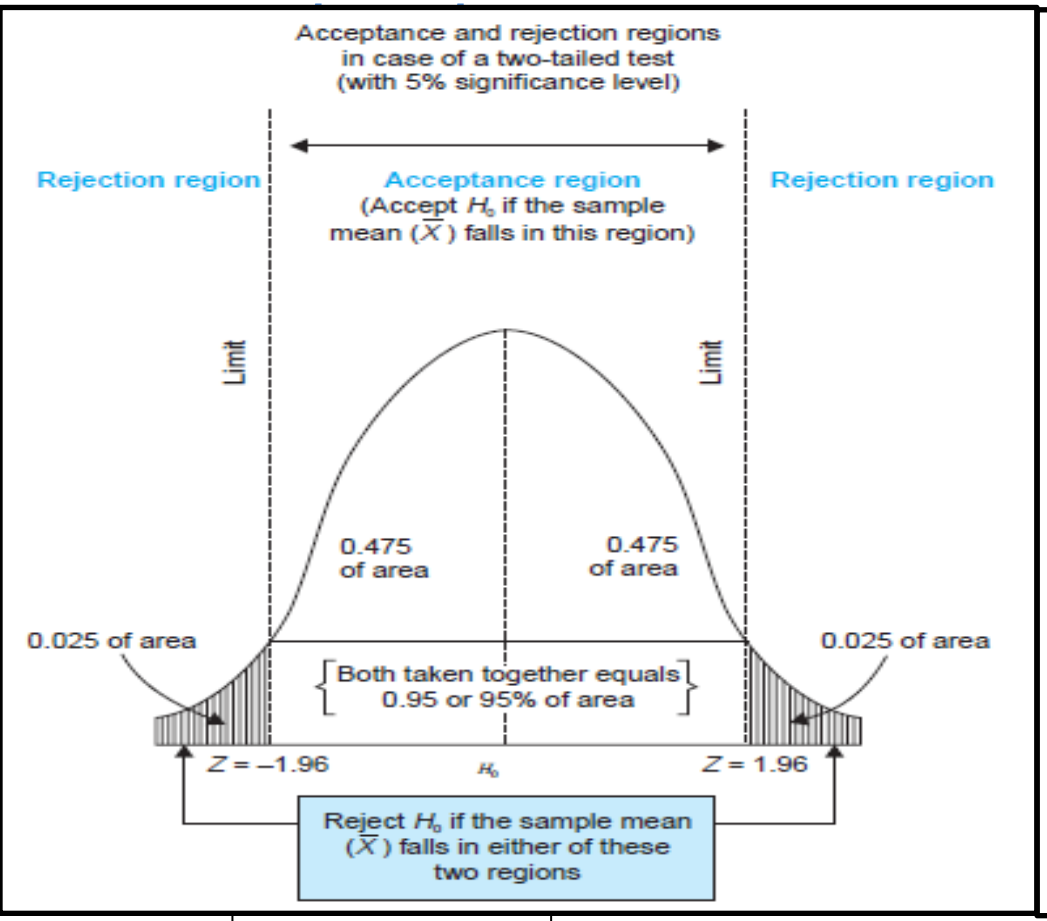
$H_0:\mu >= \mu_{H0}$ Aganist $H_a:\mu < \mu_{H0}$

If we have ≠ in alternate hypotheses – Two tailed test

If we have > sign in alternate hypotheses – **right tailed**

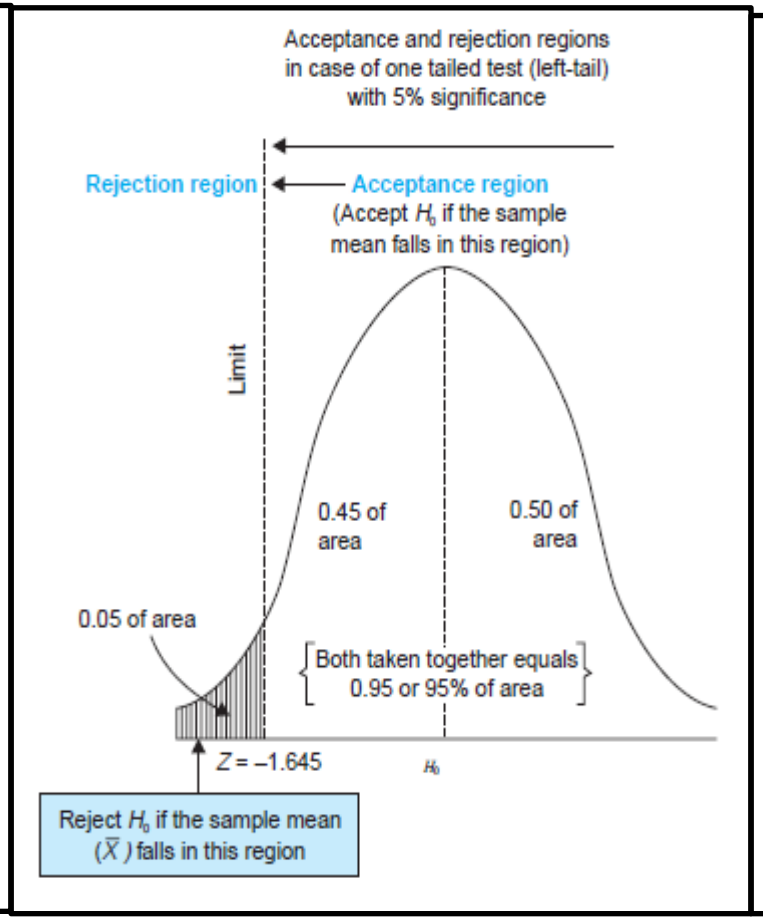
If we have < sign in alternate hypotheses – left **tailed**

One tailed and two tailed



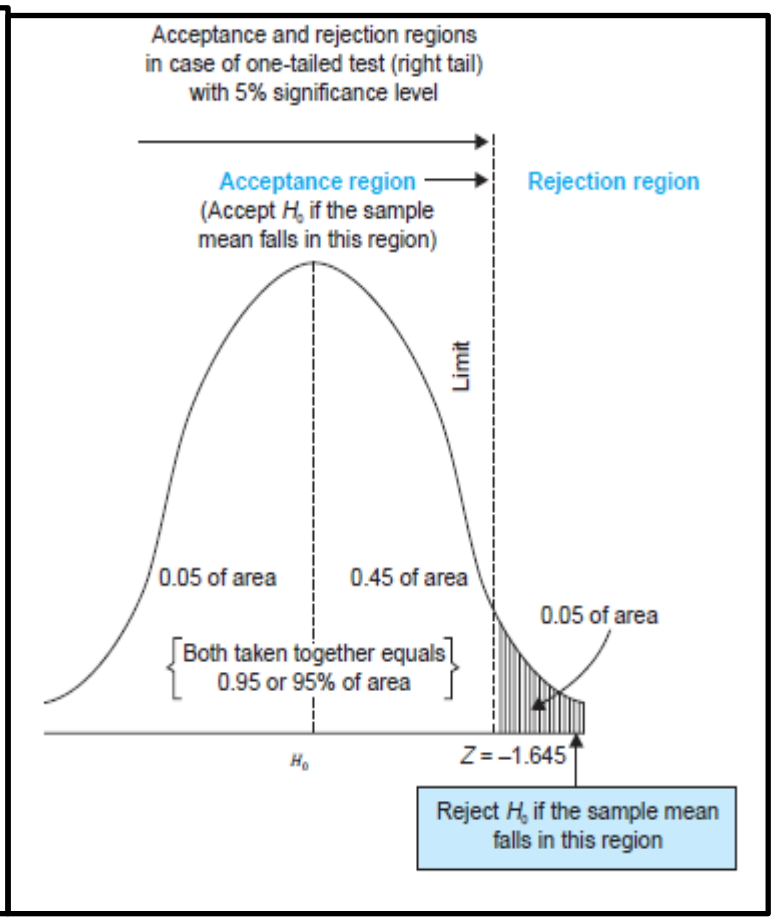
$H_0: \mu = \mu_{H0}$

$H_a: \mu \neq \mu_{H0}$



$H_0: \mu = \mu_{H0}$

$H_a: \mu < \mu_{H0}$



$H_0: \mu = \mu_{H0}$

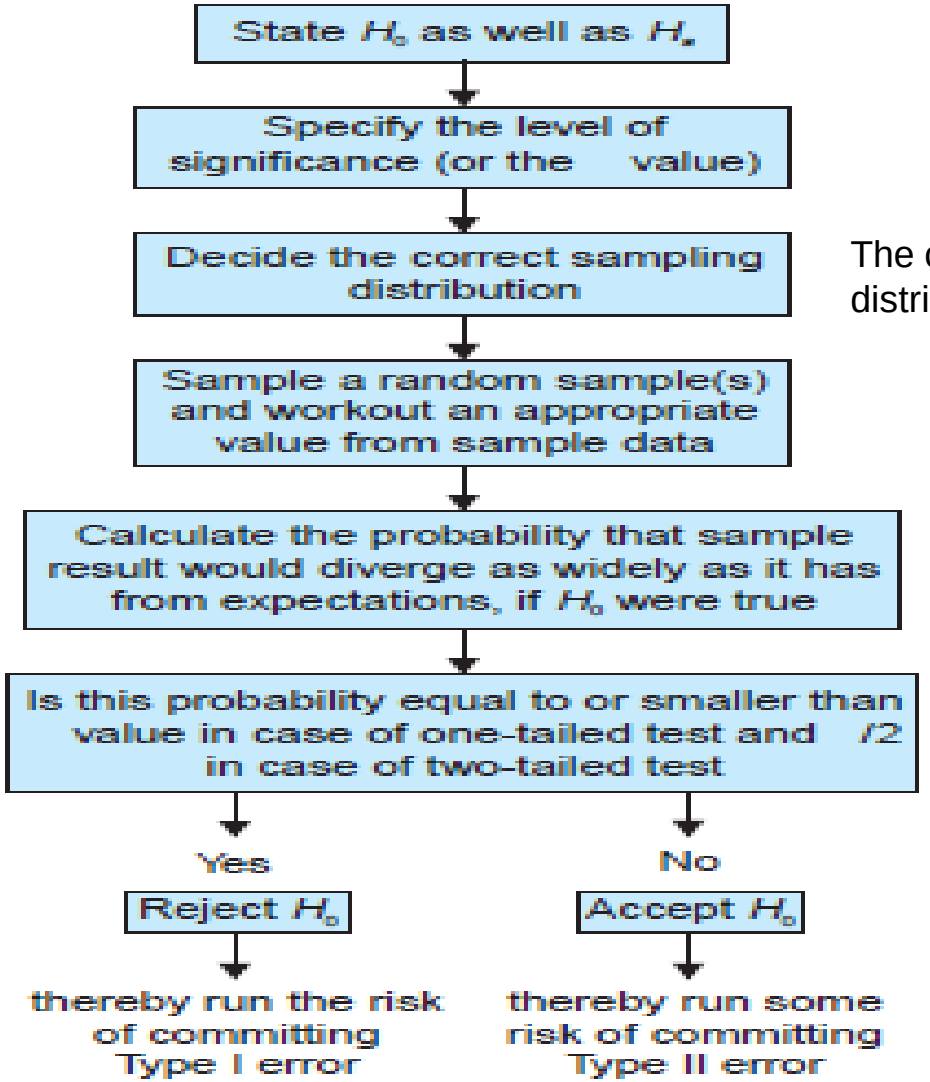
$H_a: \mu > \mu_{H0}$

One tailed and two tailed test

It should always be remembered that accepting H0 on the basis of sample information does not constitute the proof that H0 is true. We only mean that there is no statistical evidence to reject it, but we are certainly not saying that H0 is true (although we behave as if H0 is true).

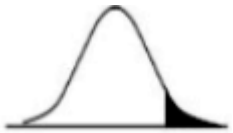
Steps in Hypothesis Testing

The factors that affect the level of significance are: (a) the magnitude of the difference between sample means; (b) the size of the samples; (c) the variability of measurements within samples



The choice generally remains between normal distribution and the t-distribution.

Type of Hypothesis Tests:

Sl No	Type of Hypothesis Test	Alternate Hypothesis	NullHypothesis
1	**Left Tailed Test** 	$<$	$\geq$
2	**Right Tailed Test** 	$>$	$\leq$
3	**Two Tailed Test** 	$\neq$	$=$

Large-Sample Tests for a Population Mean

Different Hypothesis Tests

Hypothesis Test	Test Statistic
z-test	z- statistic
t-test	t- statistic
Chi square test	Chi square statistic

Large-Sample Tests for a Population Mean

Note:

Hypothesis Test	Population S.D. known	Population S.D. unknown
$n < 30$ Small sample drawn from Normal population	Use z- test	Use t-test
$n \geq 30$ Large sample	Use z- test	Use z- test Or t- test

Example:

- **The article "Wear in Boundary Lubrication" (S. Hsu, R. Munro, and M. Shen, *Journal of Engineering Tribology*, $2002: 427-441$) discusses several experiments involving various lubricants.**

- **In one experiment, 45 steel balls lubricated with purified paraffin were subjected to a $40\ kg$ load at $600\ rpm$ for $60\ minutes$.**

Example:

- The average wear, measured by the reduction in diameter, was $673.2\ \mu m$, and the standard deviation was $14.9\ \mu m$.
- Assume that the specification for a lubricant is that the mean wear be less than $675\ \mu m$.
- Find the P-value for testing $H_0: \mu \geq 675$ versus $H_1: \mu < 675$.

Solution:

Solution:

Solution:

- **The null hypothesis is that the lubricant does not meet the specification, and that the difference between the sample mean of 673.2 and 675 is due to chance.**

- **The alternate hypothesis is that the lubricant does indeed meet the specification.**

$$z = \frac{673.2 - 675}{2.22} = -0.81$$

Soution:

- The $P-v$alue is 0.209 .
- Therefore if H_0 is true, there is a 20.9% chance to observe a sample whose disagreement with H_0 is as least as great as that which was actually observed.
- Since 0.209 is not a very small probability,
- we do not reject H_0.
- Instead, we conclude that H_0 is plausible.

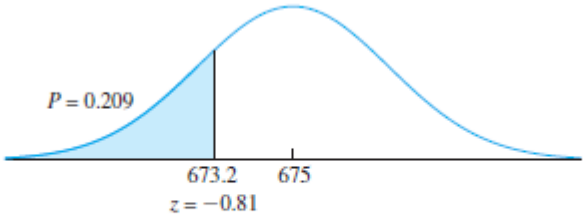


FIGURE 6.2 The null distribution of $\bar{X}$ is $N(675, 2.22^2)$. Thus if H_0 is true, the probability that $\bar{X}$ takes on a value as extreme as or more extreme than the observed value of 673.2 is 0.209. This is the P-value.

Example:

- **A scale is to be calibrated by weighing a 1000 g test weight 60 times.**

- **The 60 scale readings have mean 1000.6 g and standard deviation 2 g.**

- **Find the *P*-value for testing**

$$H_0 : \mu = 1000 \; versus \; H_1 : \mu \neq 1000.$$

Solution:

$$H_0 : \mu = 1000 \; versus \; H_1 : \mu \neq 1000.$$

We assume H_0 is true

$$z = \frac{1000.6 - 1000}{0.258}$$

$$= 2.32$$

Solution:

- **The *P*-value is the sum of the areas in both of these tails, which is 0.0204.**

- **Therefore, if H_0 is true, the probability of a result as extreme as or more extreme than that observed is only 0.0204.**

- **The evidence against H_0 is pretty strong. It would be prudent to reject H_0 and to recalibrate the scale.**

Example:

- **A sample of 400 male students is found to have a mean height 67.47 inches.**

- **Can it be reasonably regarded as a sample from a large population with mean height 67.39 inches and standard deviation1.30 inches?** Test at 5% level of significance.

Solution:

H_0: $\mu = 67.39$inches

H_1: $\mu \neq 67.39$inches

$$z = \frac{X - \mu 0}{\sigma / \sqrt{n}} = \frac{67.47 - 67.39}{130 / \sqrt{400}} = 1.231$$

H0 is accepted.

Drawing Conclusions from the Results of Hypothesis Tests

- The only two conclusions that can be reached in a hypothesis test are that
- H_0 is false or that H_0 is plausible.
- One can never conclude that H_0 is true.

Drawing Conclusions from the Results of Hypothesis Tests

- How do we know when to reject H_0?
- The smaller the P-value, the less plausible H_0 becomes.
- A common rule of thumb is to draw the line at 5%. According to this rule of thumb, if $P \leq 0.05$, H_0 is rejected; otherwise H_0 is not rejected.

Drawing Conclusions from the Results of Hypothesis Tests

- The smaller the P-value, the more certain we can be that H_0 is false.

- The larger the P-value, the more plausible H_0 becomes, but we can never be certain that H_0 is true.

Drawing Conclusions from the Results of Hypothesis Tests

- There is no sharp dividing line between conclusive evidence against H_0

- So while this rule of thumb is convenient, it has no real scientific justification.

Drawing Conclusions from the Results of Hypothesis Tests

- A rule of thumb suggests to reject H_0 whenever $P \leq 0.05$.

Drawing Conclusions from the Results of Hypothesis Tests

Statistical Significance:

- Whenever the *P*-value is less than a particular threshold, the result is said to be "statistically significant" at that level.
- So, for example, if $P \leq 0.05$, the result is statistically significant at the 5% level; if $P \leq 0.01$, the result is statistically significant at the 1% level, and so on.

Statistical Significance:

- If a result is statistically significant at the $100\alpha\%$ level, we can also say that the null hypothesis is "rejected at level $100\alpha\%$."

Drawing Conclusions from the Results of Hypothesis Tests

- The null hypothesis is rejected at the $100\alpha\%$ level.
- When reporting the result of a hypothesis test, report the P $-$value, rather than just comparing it to 5% or 1%.

Drawing Conclusions from the Results of Hypothesis Tests

- Let α be any value between 0 and 1. Then, if $P \leq \alpha$,
- The result of the test is said to be statistically significant at the $100\alpha\%$ level.

Drawing Conclusions from the Results of Hypothesis Tests

Example:

- **A hypothesis test is performed of the null hypothesis $H_0: \mu = 0$. The $P-$value turns out to be 0.03.**
- **Is the result statistically significant at the 10% level? The 5% level? The 1% level?**
- **Is the null hypothesis rejected at the 10%level? The 5% level? The1% level?**

Solution:

- **The result is statistically significant at any level greater than or equal to 3%.**
- **Thus it is statistically significant at the 10% and 5% levels, but not at the 1% level.**
- **Similarly, we can reject the null hypothesis at any level greater than or equal to 3%**
- **So H_0 is rejected at the 10% and 5% levels, but not at the 1% level.**

The *P*-value Is Not the Probability That *H*0 Is True

It makes sense to define the P-value as the probability of observing an extreme value of a statistic such as X, since the value of X could come out differently if the experiment were repeated. The null hypothesis, on the other hand, either is true or is not true. The truth or falsehood of H0 cannot be changed by repeating the experiment. It is therefore not correct to discuss the “probability” that H0 is true.

THANK YOU

Dr. Roopa Ravish
Department of Computer Science & Engineering

RESEARCH METHODOLOGY

Unit-03:
Testing of hypotheses

Dr.Roopa Ravish
Department of CSE

RESEARCH METHODOLOGY

Topic: Basic concepts - Procedure for hypothesis testing, flow diagram for hypothesis testing

Tests of Hypothesis

(a) Parametric tests or standard tests of hypotheses

(b) Non-parametric tests or distribution-free test of hypotheses.

Tests of Hypothesis

- Parametric tests usually assume certain properties of the parent population from which we draw samples.

- Assumptions like observations come from a normal population, sample size is large, assumptions about the population parameters like mean, variance, etc., must hold good before parametric tests can be used.

- But there are situations when the researcher cannot or does not want to make such assumptions. In such situations we use statistical methods for testing hypotheses which are called non-parametric tests.

- Besides, most non-parametric tests assume only nominal or ordinal data, whereas parametric tests require measurement equivalent to at least an interval scale.
- Non-parametric tests need more observations than parametric tests to achieve the same size of Type I and Type II errors.

z-test vs t-test

1. *Population normal, population infinite, sample size may be large or small but variance of the population is known, H_a may be one-sided or two-sided:*

   In such a situation z-test is used for testing hypothesis of mean and the test statistic z is worked our as under:

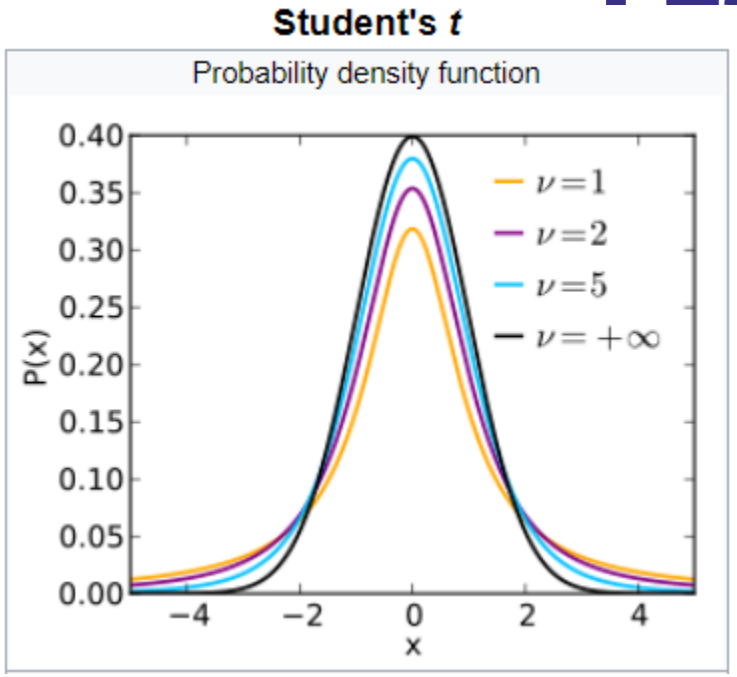
$$z = \frac{\bar{X} - \mu_{H_0}}{\sigma_p / \sqrt{n}}$$

3. *Population normal, population infinite, sample size small and variance of the population unknown, H_a may be one-sided or two-sided:*

   In such a situation t-test is used and the test statistic t is worked out as under:

$$t = \frac{\bar{X} - \mu_{H_0}}{\sigma_s / \sqrt{n}} \quad \text{with d.f.} = (n-1)$$

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



Degrees of Freedom (*df*)	Critical Value for Significance Level (Two-Tailed)			
	10%	*5%*	*1%*	*.1%*
4†	2.13	2.78	4.60	8.61
5	2.02	2.57	4.03	6.87
9†	1.83	2.26	3.25	4.78
120	1.66	1.98	2.62	3.37
1,000	1.65	1.96	2.58	3.30
Normal (*Z*)	1.64	1.96	2.58	3.29

Eg: t-test

The specimen of copper wires drawn form a large lot have the following breaking strength (in kg. weight):

578, 572, 570, 568, 572, 578, 570, 572, 596, 544

Test (using Student's *t*-statistic)whether the mean breaking strength of the lot may be taken to be 578 kg. weight (Test at 5 per cent level of significance).

$$t = \frac{\bar{X} - \mu_{H_0}}{\sigma_s/\sqrt{n}} \text{ with d.f.} = (n-1)$$

$$\sigma_s = \sqrt{\frac{\Sigma\left(X_i - \bar{X}\right)^2}{(n-1)}}$$

Eg: t-test

The specimen of copper wires drawn form a large lot have the following breaking strength (in kg. weight):

578, 572, 570, 568, 572, 578, 570, 572, 596, 544

Test (using Student's *t*-statistic)whether the mean breaking strength of the lot may be taken to be 578 kg. weight (Test at 5 per cent level of significance).

$$t = \frac{\bar{X} - \mu_{H_0}}{\sigma_s/\sqrt{n}} \text{ with d.f.} = (n-1)$$

$$\sigma_s = \sqrt{\frac{\Sigma\left(X_i - \bar{X}\right)^2}{(n-1)}}$$

Solution: Taking the null hypothesis that the population mean is equal to hypothesised mean of 578 kg., we can write:

$$H_0 : \mu = \mu_{H_0} = 578 \text{ kg.}$$

$$H_a : \mu \neq \mu_{H_0}$$

As the sample size is mall (since $n = 10$) and the population standard deviation is not known, we shall use *t*-test assuming normal population and shall work out the test statistic *t* as under:

$$t = \frac{\bar{X} - \mu_{H_0}}{\sigma_s/\sqrt{n}}$$

To find $\bar{X}$ and σ_s we make the following computations:

Eg: t-test

S. No.	X_i	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$
1	578	6	36
2	572	0	0
3	570	−2	4
4	568	−4	16
5	572	0	0
6	578	6	36
7	570	−2	4
8	572	0	0
9	596	24	576
10	544	−28	784
$n = 10$	$\sum X_i = 5720$		$\sum (X_i - \bar{X})^2 = 1456$

Eg: t-test

$$\therefore \quad \bar{X} = \frac{\Sigma X_i}{n} = \frac{5720}{10} = 572 \text{ kg.}$$

and $$\sigma_s = \sqrt{\frac{\Sigma\left(X_i - \bar{X}\right)^2}{n-1}} = \sqrt{\frac{1456}{10-1}} = 12.72 \text{ kg.}$$

Hence, $$t = \frac{572 - 578}{12.72/\sqrt{10}} = -1.488$$

Degree of freedom = $(n - 1) = (10 - 1) = 9$

As H_a is two-sided, we shall determine the rejection region applying two-tailed test at 5 per cent level of significance, and it comes to as under, using table of t-distribution* for 9 d.f.:

$$R : |t| > 2.262$$

As the observed value of t (i.e., – 1.488) is in the acceptance region, we accept H_0 at 5 per cent level and conclude that the mean breaking strength of copper wires lot may be taken as 578 kg. weight.

Eg:

The mean of a certain production process is known to be 50 with a standard deviation of 2.5. The production manager may welcome any change is mean value towards higher side but would like to safeguard against decreasing values of mean. He takes a sample of 12 items that gives a mean value of 48.5. What inference should the manager take for the production process on the basis of sample results? Use 5 per cent level of significance for the purpose.

Eg:

The mean of a certain production process is known to be 50 with a standard deviation of 2.5. The production manager may welcome any change is mean value towards higher side but would like to safeguard against decreasing values of mean. He takes a sample of 12 items that gives a mean value of 48.5. What inference should the manager take for the production process on the basis of sample results? Use 5 per cent level of significance for the purpose.

Solution: Taking the mean value of the population to be 50, we may write:

$$H_0 : \mu_{H_0} = 50$$

$H_a : \mu_{H_0} < 50$ (Since the manager wants to safeguard against decreasing values of mean.)

Eg:

and the given information as $X = 48.5$, $\sigma_p = 2.5$ and $n = 12$. Assuming the population to be normal, we can work out the test statistic z as under:

$$z = \frac{\bar{X} - \mu_{H_0}}{\sigma_p/\sqrt{n}} = \frac{48.5 - 50}{2.5/\sqrt{12}} = -\frac{1.5}{(2.5)/(3.464)} = -2.0784$$

As H_a is one-sided in the given question, we shall determine the rejection region applying one-tailed test (in the left tail because H_a is of less than type) at 5 per cent level of significance and it comes to as under, using normal curve area table:

$$R : z < -1.645$$

The observed value of z is -2.0784 which is in the rejection region and thus, H_0 is rejected at 5 per cent level of significance. We can conclude that the production process is showing mean which is significantly less than the population mean and this calls for some corrective action concerning the said process.

Chi-Square tests χ^2

A chi-square goodness of fit test determines if a sample data matches a population.

Used to obtain confidence interval estimate of unknown population variance.
Non-parametric test and as such no rigid assumptions are necessary in respect of type of population.

chi-square can be used (i) as a test of goodness of fit and (ii) as a test of independence.

As a test of goodness of fit, test enables us to see how well does the assumed theoretical distribution (such as Binomial distribution, Poisson distribution or Normal distribution) fit to the observed data.
If the calculated value of χ^2 is less than the table value at a certain level of significance, the fit is considered to be a good one which means that the divergence between the observed and expected frequencies is attributable to fluctuations of sampling. But if the calculated value of χ^2 is greater than its table value, the fit is not considered to be a good one.

Chi-Square tests

As a test of independence, χ^2 test enables us to explain whether or not two attributes are associated (Independent Variable/Dependent Variable).

On this basis we first calculate the expected frequencies and then work out the value of χ^2 . If the calculated value of χ^2 is less than the table value at a certain level of significance for given degrees of freedom, we conclude that null hypothesis stands which means that the two attributes are independent or not associated .

But if the calculated value of χ^2 is greater than its table value, our inference then would be that null hypothesis does not hold good which means the two attributes are associated and the association is not because of some chance factor but it exists in reality. It may, however, be stated here that χ^2 is not a measure of the degree of relationship or the form of relationship between two attributes, but is simply a technique of judging the significance of such association or relationship between two attributes.

Conditions for chi-square test.

The following conditions should be satisfied before χ^2 test can be applied:

(i) Observations recorded and used are collected on a random basis.

(ii) All the itmes in the sample must be independent.

(iii) No group should contain very few items, say less than 10. In case where the frequencies are less than 10, regrouping is done by combining the frequencies of adjoining groups so that the new frequencies become greater than 10. Some statisticians take this number as 5, but 10 is regarded as better by most of the statisticians.

(iv) The overall number of items must also be reasonably large. It should normally be at least 50, howsoever small the number of groups may be.

(v) The constraints must be linear. Constraints which involve linear equations in the cell frequencies of a contingency table (i.e., equations containing no squares or higher powers of the frequencies) are known are know as linear constraints.

STEPS INVOLVED IN APPLYING CHI-SQUARE TEST

(i) First of all calculate the expected frequencies on the basis of given hypothesis or on the basis of null hypothesis. Usually in case of a 2×2 or any contingency table, the expected frequency for any given cell is worked out as under:

$$\text{Expected frequency of any cell} = \left[\frac{\begin{array}{c}\text{(Row total for the row of that cell)} \times \\ \text{(Column total for the column of that cell)}\end{array}}{\text{(Grand total)}}\right]$$

(ii) Obtain the difference between observed and expected frequencies and find out the squares of such differences i.e., calculate $(O_{ij} - E_{ij})^2$.

(iii) Divide the quantity $(O_{ij} - E_{ij})^2$ obtained as stated above by the corresponding expected frequency to get $(O_{ij} - E_{ij})^2/E_{ij}$ and this should be done for all the cell frequencies or the group frequencies.

(iv) Find the summation of $(O_{ij} - E_{ij})^2/E_{ij}$ values or what we call $\sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$. This is the required χ^2 value.

The χ^2 value obtained as such should be compared with relevant table value of χ^2 and then inference be drawn as stated above.

Example

A die is thrown 132 times with following results:

Number turned up	1	2	3	4	5	6
Frequency	16	20	25	14	29	28

Is the die unbiased?

Example

A die is thrown 132 times with following results:

Number turned up	1	2	3	4	5	6
Frequency	16	20	25	14	29	28

Is the die unbiased?

Solution: Let us take the hypothesis that the die is unbiased. If that is so, the probability of obtaining any one of the six numbers is 1/6 and as such the expected frequency of any one number coming upward is $132 \times 1/6 = 22$. Now we can write the observed frequencies along with expected frequencies and work out the value of χ^2 as follows:

No. turned up	*Observed frequency* O_i	*Expected frequency* E_i	$(O_i - E_i)$	$(O_i - E_i)^2$	$(O_i - E_i)^2/E_i$
1	16	22	–6	36	36/22
2	20	22	–2	4	4/22
3	25	22	3	9	9/22
4	14	22	–8	64	64/22
5	29	22	7	49	49/22
6	28	22	6	36	36/22

Example

$$\therefore \qquad \sum [(O_i - E_i)^2/E_i] = 9.$$

Hence, the calculated value of $\chi^2 = 9$.

$\because$ Degrees of freedom in the given problem is

$(n - 1) = (6 - 1) = 5$.

The table value* of χ^2 for 5 degrees of freedom at 5 per cent level of significance is 11.071. Comparing calculated and table values of χ^2, we find that calculated value is less than the table value and as such could have arisen due to fluctuations of sampling. The result, thus, supports the hypothesis and it can be concluded that the die is unbiased.

Example 2

Find the value of χ^2 for the following information:

Class	*A*	*B*	*C*	*D*	*E*
Observed frequency	8	29	44	15	4
Theoretical (or expected) frequency	7	24	38	24	7

Solution: Since some of the frequencies less than 10, we shall first re-group the given data as follows and then will work out the value of χ^2 :

Example 2

Class	*Observed frequency* O_i	*Expected frequency* E_i	$O_i - E_i$	$(O_i - E_i)^2/E_i$
A and *B*	(8 + 29) = 37	(7 + 24) = 31	6	36/31
C	44	38	6	36/38
D and *E*	(15 + 4) = 19	(24 + 7) = 31	–12	144/31

$$\therefore \qquad \chi^2 = \Sigma \frac{(O_i - E_i)^2}{E_i} = 6.76 \text{ app.}$$

Example 2

Class	*Observed frequency* O_i	*Expected frequency* E_i	$O_i - E_i$	$(O_i - E_i)^2/E_i$
A and *B*	(8 + 29) = 37	(7 + 24) = 31	6	36/31
C	44	38	6	36/38
D and *E*	(15 + 4) = 19	(24 + 7) = 31	–12	144/31

$$\therefore \qquad \chi^2 = \Sigma \frac{(O_i - E_i)^2}{E_i} = 6.76 \text{ app.}$$

Problem - 3

Genetic theory states that children having one parent of blood type *A* and the other of blood type *B* will always be of one of three types, *A*, *AB*, *B* and that the proportion of three types will on an average be as 1 : 2 : 1. A report states that out of 300 children having one *A* parent and *B* parent,30 per cent were found to be types *A*, 45 per cent per cent type *AB* and remainder type *B*. Test the hypothesis by test

Class	Obs Freq	Exp Freq	Oi – Ei	(Oi – Ei)^2/Ei
A				
AB				
B				

Answer: Problem - 3

The expected frequencies of type A, AB and B (as per the genetic theory) should have been 75, 150 and 75 respectively.

We now calculate the value of χ^2 as follows:

Table 10.4

Type	*Observed frequency* O_i	*Expected frequency* E_i	$(O_i - E_i)$	$(O_i - E_i)^2$	$(O_i - E_i)^2/E_i$
A	90	75	15	225	225/75 = 3
AB	135	150	–15	225	225/150 = 1.5
B	75	75	0	0	0/75 = 0

$$\therefore \qquad \chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 3 + 1.5 + 0 = 4.5$$

$$\because \qquad \text{d.f.} = (n - 1) = (3 - 1) = 2.$$

Table value of χ^2 for 2 d.f. at 5 per cent level of significance is 5.991.

The calculated value of χ^2 is 4.5 which is less than the table value and hence can be ascribed to have taken place because of chance. This supports the theoretical hypothesis of the genetic theory that on an average type A, AB and B stand in the proportion of 1 : 2 : 1.

Problem

The table given below shows the data obtained during outbreak of smallpox:

	Attacked	*Not attacked*	*Total*
Vaccinated	31	469	500
Not vaccinated	185	1315	1500
Total	216	1784	2000

Test the effectiveness of vaccination in preventing the attack from smallpox. Test your result with the help of χ^2 at 5 per cent level of significance.

Problem

Solution: Let us take the hypothesis that vaccination is not effective in preventing the attack from smallpox i.e., vaccination and attack are independent. On the basis of this hypothesis, the expected frequency corresponding to the number of persons vaccinated and attacked would be:

$$\text{Expectation of } (AB) = \frac{(A) \times (B)}{N}$$

when A represents vaccination and B represents attack.

$\therefore$

$$(A) = 500$$
$$(B) = 216$$
$$N = 2000$$

$$\text{Expectation of } (AB) = \frac{500 \times 216}{2000} = 54$$

Now using the expectation of (AB), we can write the table of expected values as follows:

The table shows the data obtained during outbreak of smallpox.Test the effectiveness of the vaccine at 5% significance level.

H0: The vaccine has no effect; Ha: Vaccine is effective.

Ob Freq	Attacked(A)	Not Attacked(NA)	Row Tol
Vaccinated(V)	31	469	500
Not Vaccinated (NV)	185	1315	1500
Col Total	216	1784	2000

Exp Freq	Attacked	Not Attacked	Row Tol
Vaccinated	500*216/2000 =54	446	500
Not Vaccinated	162	1500*1784/2000= 1338	1500
Col Total	216	1784	2000

Class	Obs Freq	Exp Freq	Oi – Ei	(Oi – Ei)^2/Ei
V-A	31	54	-23	-23^2/54=9.80
V-NA	469	446	23	23^2/446=1.19
NV-A	185	162	23	23^2/162=3.27
NV-NA	1315	1338	-23	23^2/1338=0.40

$$\chi^2 = \Sigma \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} = 14.642$$

$\because$ Degrees of freedom in this case = $(r - 1)\,(c - 1) = (2 - 1)\,(2 - 1) = 1$.

The table value of χ^2 for 1 degree of freedom at 5 per cent level of significance is 3.841. The calculated value of χ^2 is much higher than this table value and hence the result of the experiment does not support the hypothesis. We can, thus, conclude that vaccination is effective in preventing the attack from smallpox.

Home work -1

Two research workers classified some people in income groups on the basis of sampling studies. Their results are as follows:

Investigators	*Income groups*			*Total*
	Poor	*Middle*	*Rich*	
A	160	30	10	200
B	140	120	40	300
Total	300	150	50	500

Home work

3. An experiment was conducted to test the efficacy of chloromycetin in checking typhoid. In a certain hospital chloromycetin was given to 285 out of the 392 patients suffering from typhoid. The number of typhoid cases were as follows:

	Typhoid	*No Typhoid*	*Total*
Chloromycetin	35	250	285
No chloromycetin	50	57	107
Total	85	307	392

With the help of χ^2, test the effectiveness of chloromycetin in checking typhoid.

(The χ^2 value at 5 per cent level of significance for one degree of freedom is 3.841).

THANK YOU

Dr. Roopa Ravish
Department of Computer Science & Engineering

RESEARCH METHODOLOGY

Topic: Basic concepts - ANALYSIS OF VARIANCE (ANOVA)

Dr. Roopa Ravish
Department of Computer Science & Engineering

ANOVA: Analysis of Variance

- Used of hypothesis testing when >2 population/samples cases are involved

1. *Population normal, population infinite, sample size may be large or small but variance of the population is known, H_a may be one-sided or two-sided:*

In such a situation z-test is used for testing hypothesis of mean and the test statistic z is worked our as under:

$$z = \frac{\overline{X} - \mu_{H_0}}{\sigma_p / \sqrt{n}}$$

3. *Population normal, population infinite, sample size small and variance of the population unknown, H_a may be one-sided or two-sided:*

In such a situation t-test is used and the test statistic t is worked out as under:

$$t = \frac{\overline{X} - \mu_{H_0}}{\sigma_s / \sqrt{n}} \text{ with d.f.} = (n-1)$$

$$\sigma_s = \sqrt{\frac{\Sigma\left(X_i - \overline{X}\right)^2}{(n-1)}}$$

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\Sigma\left(X_{1i} - \overline{X}_1\right)^2 + \Sigma\left(X_{2i} - \overline{X}_2\right)^2}{n_1 + n_2 - 2}} \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

ANOVA: Analysis of Variance

- This technique is used when multiple sample cases are involved.

- The significance of the difference between the means of two samples can be judged through either *z*-test or the *t*-test, but the difficulty arises when we happen to examine the significance
of the difference amongst more than two sample means at the same time.

- The ANOVA technique enables us to perform this simultaneous test and as such is considered to be an important tool of analysis in the hands of a researcher. Using this technique, one can draw inferences about whether the samples have been drawn from populations having the same mean.

ANOVA: Analysis of Variance

- The ANOVA technique is important in the context of all those situations where we want to compare more than two populations such as in comparing the yield of crop from several varieties of seeds, the gasoline mileage of four automobiles, the smoking habits of five groups of university students and so on.

- Therefore, one quite often utilizes the ANOVA technique and through it investigates the differences among the means of all the populations simultaneously.

- "The essence of ANOVA is that the total amount of variation in a set of data is broken down into two types, that amount which can be attributed to chance and that amount which can be attributed to specified causes."

- There may be variation between samples and also within sample items. ANOVA consists in splitting the variance for analytical purposes.

ANOVA: Analysis of Variance

Hence, it is a method of analysing the variance to which a response is subject into its various components corresponding to various sources of variation.

Through this technique one can explain whether various varieties of seeds or fertilizers or soils differ significantly so that a policy decision could be taken accordingly, concerning a particular variety in the context of agriculture researches.

Thus, through ANOVA technique one can, in general, investigate any number of factors which are hypothesized or said to influence the dependent variable.

If we take only one factor and investigate the differences amongst its various categories having numerous possible values, we are said to use one-way ANOVA and in case we investigate two factors at the same time, then we use two-way ANOVA. In a two or more way ANOVA, the interaction (i.e., inter-relation between two independent variables/factors), if any, between two independent variables affecting a dependent variable can as well be studied for better decisions.

ANOVA: Analysis of Variance

THE BASIC PRINCIPLE OF ANOVA

we have to make two estimates of population variance viz., one based on between samples variance and the other based on within samples variance.
Then the said two estimates of population variance are compared with *F*-test, wherein we work out.

$$F = \frac{\text{Estimate of population variance based on between samples variance}}{\text{Estimate of population variance based on within samples variance}}$$

This value of *F* is to be compared to the *F*-limit for given degrees of freedom.
If the *F* value we work out is equal or exceeds* the *F*-limit value we may say that there are significant differences between the sample means.

ANOVA TECHNIQUE: One Way (Single Factor)

One-way (or single factor) ANOVA: Under the one-way ANOVA, we consider only one factor and then observe that the reason for said factor to be important is that several possible types of samples can occur within that factor. We then determine if there are differences within that factor. The technique involves the following steps:

(i) Obtain the mean of each sample i.e., obtain

$$\bar{X}_1, \bar{X}_2, \bar{X}_3, \ldots, \bar{X}_k$$

when there are k samples.

(ii) Work out the mean of the sample means as follows:

$$\overline{\overline{X}} = \frac{\bar{X}_1 + \bar{X}_2 + \bar{X}_3 + \ldots + \bar{X}_k}{\text{No. of samples } (k)}$$

(iii) Take the deviations of the sample means from the mean of the sample means and calculate the square of such deviations which may be multiplied by the number of items in the corresponding sample, and then obtain their total. This is known as the sum of squares for variance between the samples (or *SS* between). Symbolically, this can be written:

$$SS \text{ between} = n_1\left(\bar{X}_1 - \overline{\overline{X}}\right)^2 + n_2\left(\bar{X}_2 - \overline{\overline{X}}\right)^2 + \ldots + n_k\left(\bar{X}_k - \overline{\overline{X}}\right)^2$$

ANOVA TECHNIQUE: One Way (Single Factor)

(iv) Divide the result of the (iii) step by the degrees of freedom between the samples to obtain variance or mean square (*MS*) between samples. Symbolically, this can be written:

$$MS \text{ between} = \frac{SS \text{ between}}{(k-1)}$$

where $(k-1)$ represents degrees of freedom (d.f.) between samples.

(v) Obtain the deviations of the values of the sample items for all the samples from corresponding means of the samples and calculate the squares of such deviations and then obtain their total. This total is known as the sum of squares for variance within samples (or *SS* within). Symbolically this can be written:

$$SS \text{ within} = \Sigma\left(X_{1i} - \overline{X}_1\right)^2 + \Sigma\left(X_{2i} - \overline{X}_2\right)^2 + \ldots + \Sigma\left(X_{ki} - \overline{X}_k\right)^2$$

$$i = 1, 2, 3, \ldots$$

(vi) Divide the result of (v) step by the degrees of freedom within samples to obtain the variance or mean square (*MS*) within samples. Symbolically, this can be written:

ANOVA TECHNIQUE: One Way (Single Factor)

$$MS \text{ within} = \frac{SS \text{ within}}{(n-k)}$$

where $(n-k)$ represents degrees of freedom within samples,

n = total number of items in all the samples i.e., $n_1 + n_2 + \ldots + n_k$

k = number of samples.

(vii) For a check, the sum of squares of deviations for total variance can also be worked out by adding the squares of deviations when the deviations for the individual items in all the samples have been taken from the mean of the sample means. Symbolically, this can be written:

$$SS \text{ for total variance } = \sum\left(X_{ij} - \overline{\overline{X}}\right)^2 \qquad i = 1, 2, 3, \ldots$$

$$j = 1, 2, 3, \ldots$$

This total should be equal to the total of the result of the (iii) and (v) steps explained above i.e.,

SS for total variance = SS between + SS within.

The degrees of freedom for total variance will be equal to the number of items in all samples minus one i.e., $(n-1)$. The degrees of freedom for between and within must add up to the degrees of freedom for total variance i.e.,

$$(n-1) = (k-1) + (n-k)$$

This fact explains the additive property of the ANOVA technique.

ANOVA TECHNIQUE: One Way (Single Factor)

(viii) Finally, F-ratio may be worked out as under:

$$F\text{-ratio} = \frac{MS \text{ between}}{MS \text{ within}}$$

This ratio is used to judge whether the difference among several sample means is significant or is just a matter of sampling fluctuations. For this purpose we look into the table*, giving the values of F for given degrees of freedom at different levels of significance. If the worked out value of F, as stated above, is less than the table value of F, the difference is taken as insignificant i.e., due to chance and the null-hypothesis of no difference between sample means stands. In case the calculated value of F happens to be either equal or more than its table value, the difference is considered as significant (which means the samples could not have come from the same universe) and accordingly the conclusion may be drawn. The higher the calculated value of F is above the table value, the more definite and sure one can be about his conclusions.

ANOVA TECHNIQUE: One Way (Single Factor)

Illustration 1

Set up an analysis of variance table for the following per acre production data for three varieties of wheat, each grown on 4 plots and state if the variety differences are significant.

Plot of land	*Per acre production data*		
	Variety of wheat		
	A	*B*	*C*
1	6	5	5
2	7	5	4
3	3	3	3
4	8	7	4

ANOVA TECHNIQUE: One Way (Single Factor)

Solution through direct method: First we calculate the mean of each of these samples:

$$\bar{X}_1 = \frac{6+7+3+8}{4} = 6$$

$$\bar{X}_2 = \frac{5+5+3+7}{4} = 5$$

$$\bar{X}_3 = \frac{5+4+3+4}{4} = 4$$

Mean of the sample means or $\bar{\bar{X}} = \frac{\bar{X}_1 + \bar{X}_2 + \bar{X}_3}{k}$

$$= \frac{6+5+4}{3} = 5$$

ANOVA TECHNIQUE: One Way (Single Factor)

Now we work out *SS* between and *SS* within samples:

$$SS \text{ between} = n_1\left(\bar{X}_1 - \bar{\bar{X}}\right)^2 + n_2\left(\bar{X}_2 - \bar{\bar{X}}\right)^2 + n_3\left(\bar{X}_3 - \bar{\bar{X}}\right)^2$$

$$= 4(6-5)^2 + 4(5-5)^2 + 4(4-5)^2$$
$$= 4 + 0 + 4$$
$$= 8$$

$$SS \text{ within} = \Sigma\left(X_{1i} - \bar{X}_1\right)^2 + \Sigma\left(X_{2i} - \bar{X}_2\right)^2 + \Sigma\left(X_{3i} - \bar{X}_3\right)^2, \qquad i = 1, 2, 3, 4$$

$$= \{(6-6)^2 + (7-6)^2 + (3-6)^2 + (8-6)^2\}$$
$$+ \{(5-5)^2 + (5-5)^2 + (3-5)^2 + (7-5)^2\}$$
$$+ \{(5-4)^2 + (4-4)^2 + (3-4)^2 + (4-4)^2\}$$
$$= \{0 + 1 + 9 + 4\} + \{0 + 0 + 4 + 4\} + \{1 + 0 + 1 + 0\}$$
$$= 14 + 8 + 2$$
$$= 24$$

ANOVA TECHNIQUE: One Way (Single Factor)

$$SS \text{ for total variance } = \sum\left(X_{ij} - \overline{\overline{X}}\right)^2 \qquad i = 1, 2, 3\ldots$$

$$j = 1, 2, 3\ldots$$

$$= (6-5)^2 + (7-5)^2 + (3-5)^2 + (8-5)^2$$
$$+ (5-5)^2 + (5-5)^2 + (3-5)^2$$
$$+ (7-5)^2 + (5-5)^2 + (4-5)^2$$
$$+ (3-5)^2 + (4-5)^2$$
$$= 1 + 4 + 4 + 9 + 0 + 0 + 4 + 4 + 0 + 1 + 4 + 1$$
$$= 32$$

Alternatively, it (*SS* for total variance) can also be worked out thus:

SS for total = *SS* between + *SS* within

$$= 8 + 24$$
$$= 32$$

ANOVA TECHNIQUE: One Way (Single Factor)

We can now set up the ANOVA table for this problem:

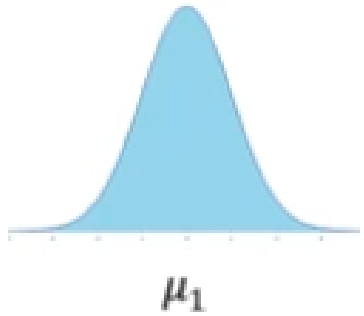
Table 11.2

Source of variation	*SS*	*d.f.*	*MS*	*F-ratio*	*5% F-limit (from the F-table)*
Between sample	8	$(3-1)=2$	$8/2=4.00$	$4.00/2.67=1.5$	$F(2,9)=4.26$
Within sample	24	$(12-3)=9$	$24/9=2.67$		
Total	32	$(12-1)=11$			

The above table shows that the calculated value of F is 1.5 which is less than the table value of 4.26 at 5% level with d.f. being $v_1 = 2$ and $v_2 = 9$ and hence could have arisen due to chance. This analysis supports the null-hypothesis of no difference is sample means. We may, therefore, conclude that the difference in wheat output due to varieties is insignificant and is just a matter of chance.

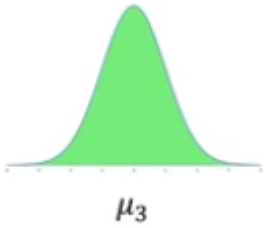
ANOVA

Compare 3 population means to see if they are different



Is one mean so far away , it is from a different population

Do all the 3 means come from the same population



Do all of these come from different population

Per Acre yeild			
Plot of land	Variety of Wheat		
	A	B	C
1	6	5	5
2	7	5	4
3	3	3	3
4	8	7	4

ANOVA

Per Acre yield			
Plot of land	Variety of Wheat		
	A	B	C
1	6	5	5
2	7	5	4
3	3	3	3
4	8	7	4
$\bar{x}$	6	5	4

$$\bar{X}_1, \bar{X}_2, \bar{X}_3, \ldots, \bar{X}_k \qquad \bar{\bar{X}} = \frac{\bar{X}_1 + \bar{X}_2 + \bar{X}_3 + \ldots + \bar{X}_k}{\text{No. of samples } (k)}$$

$$SS \text{ between} = n_1\left(\bar{X}_1 - \bar{\bar{X}}\right)^2 + n_2\left(\bar{X}_2 - \bar{\bar{X}}\right)^2 + \ldots + n_k\left(\bar{X}_k - \bar{\bar{X}}\right)^2$$

$$MS \text{ between} = \frac{SS \text{ between}}{(k-1)}$$

$$SS \text{ within} = \sum\left(X_{1i} - \bar{X}_1\right)^2 + \sum\left(X_{2i} - \bar{X}_2\right)^2 + \ldots + \sum\left(X_{ki} - \bar{X}_k\right)^2$$

$$MS \text{ within} = \frac{SS \text{ within}}{(n-k)}$$

$$F\text{-ratio} = \frac{MS \text{ between}}{MS \text{ within}}$$

$$SS \text{ for total variance} = \sum\left(X_{ij} - \bar{\bar{X}}\right)^2$$

SS for total variance = SS between + SS within.

$(n-1) = (k-1) + (n-k)$

n = total number of items in all the samples
i.e., $n_1 + n_2 + \ldots + n_k$

k = number of samples

ANOVA

Source of Variation	Sum of Squares (SS)	Deg of Freedom	Mean Sqaure(MS)	F- Ratio
Between	SS Between	(k-1)	MS Between = SS Between/(k-1)	MS between / MS Within
Within	SS Within	(n-k)	MS Within = SS within/(n-k)	
Total	SS Total	(n -1)		

$$SS \text{ between} = n_1\left(\bar{X}_1 - \bar{\bar{X}}\right)^2 + n_2\left(\bar{X}_2 - \bar{\bar{X}}\right)^2 + \ldots + n_k\left(\bar{X}_k - \bar{\bar{X}}\right)^2$$

$$SS \text{ within} = \sum\left(X_{1i} - \bar{X}_1\right)^2 + \sum\left(X_{2i} - \bar{X}_2\right)^2 + \ldots + \sum\left(X_{ki} - \bar{X}_k\right)^2$$

$$SS \text{ for total variance} = \sum\left(X_{ij} - \bar{\bar{X}}\right)^2$$

ANOVA

Per Acre yield			
Plot of land	Variety of Wheat		
	A	B	C
1	6	5	5
2	7	5	4
3	3	3	3
4	8	7	4
$\bar{x}$	6	5	4

n = total number of items in all the sample
i.e., $n_1 + n_2 + \ldots + n_k$

k = number of samples

$$\bar{X}_1 = \frac{6+7+3+8}{4} = 6$$

$$\bar{X}_2 = \frac{5+5+3+7}{4} = 5$$

$$\bar{X}_3 = \frac{5+4+3+4}{4} = 4$$

$$\bar{\bar{X}} = \frac{\bar{X}_1 + \bar{X}_2 + \bar{X}_3}{k}$$

$$= \frac{6+5+4}{3} = 5$$

$$SS \text{ for total variance} = \sum \left(X_{ij} - \bar{\bar{X}} \right)^2 \qquad i = 1, 2, 3\ldots \quad j = 1, 2, 3\ldots$$

$$= (6-5)^2 + (7-5)^2 + (3-5)^2 + (8-5)^2 + (5-5)^2 + (5-5)^2 + (3-5)^2 + (7-5)^2 + (5-5)^2 + (4-5)^2 + (3-5)^2 + (4-5)^2$$

$$= 1+4+4+9+0+0+4+4+0+1+4+1$$

$$= 32$$

$$SS \text{ between} = n_1 \left(\bar{X}_1 - \bar{\bar{X}} \right)^2 + n_2 \left(\bar{X}_2 - \bar{\bar{X}} \right)^2 + n_3 \left(\bar{X}_3 - \bar{\bar{X}} \right)^2$$

$$= 4(6-5)^2 + 4(5-5)^2 + 4(4-5)^2$$

$$= 4 + 0 + 4$$

$$= 8$$

$$SS \text{ within} = \sum \left(X_{1i} - \bar{X}_1 \right)^2 + \sum \left(X_{2i} - \bar{X}_2 \right)^2 + \sum \left(X_{3i} - \bar{X}_3 \right)^2,$$

$$= \{(6-6)^2 + (7-6)^2 + (3-6)^2 + (8-6)^2\} + \{(5-5)^2 + (5-5)^2 + (3-5)^2 + (7-5)^2\} + \{(5-4)^2 + (4-4)^2 + (3-4)^2 + (4-4)^2\}$$

$$= \{0+1+9+4\} + \{0+0+4+4\} + \{1+0+1+0\}$$

$$= 14 + 8 + 2$$

$$= 24$$

ANOVA

Per Acre yield			
Plot of land	ariety of Wheat		
	A	B	C
1	6	5	5
2	7	5	4
3	3	3	3
4	8	7	4
$\bar{x}$	6	5	4

Source of variation	*SS*	*d.f.*	*MS*	*F-ratio*	*5% F-limit (from the F-table)*
Between sample	8	$(3-1)=2$	$8/2=4.00$	$4.00/2.67=1.5$	$F(2,9)=4.26$
Within sample	24	$(12-3)=9$	$24/9=2.67$		
Total	32	$(12-1)=11$			

n = total number of items in all the samples
i.e., $n_1 + n_2 + \ldots + n_k$

k = number of samples

ANOVA

Per Acre yield			
Plot of land	ariety of Wheat		
	A	B	C
1	6	5	5
2	7	5	4
3	3	3	3
4	8	7	4
$\bar{x}$	6	5	4

n = total number of items in all the samples
i.e., $n_1 + n_2 + \ldots + n_k$

k = number of samples

Source of variation	*SS*	*d.f.*	*MS*	*F-ratio*	*5% F-limit (from the F-table)*
Between sample	8	$(3-1)=2$	$8/2=4.00$	$4.00/2.67=1.5$	$F(2,9)=4.26$
Within sample	24	$(12-3)=9$	$24/9=2.67$		
Total	32	$(12-1)=11$			

Anova: Single Factor						
SUMMARY						
Groups	Count	Sum	Average	Variance		
A	4	24	6	4.666667		
B	4	20	5	2.666667		
C	4	16	4	0.666667		
ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	8	2	4	1.5	0.274016	4.256495
Within Groups	24	9	2.666667			
Total	32	11				

THANK YOU

Dr. Roopa Ravish
Department of Computer Science & Engineering

RESEARCH METHODOLOGY

Topic: Data Representation

Department of Computer Science & Engineering

PRESENTATION OF DATA

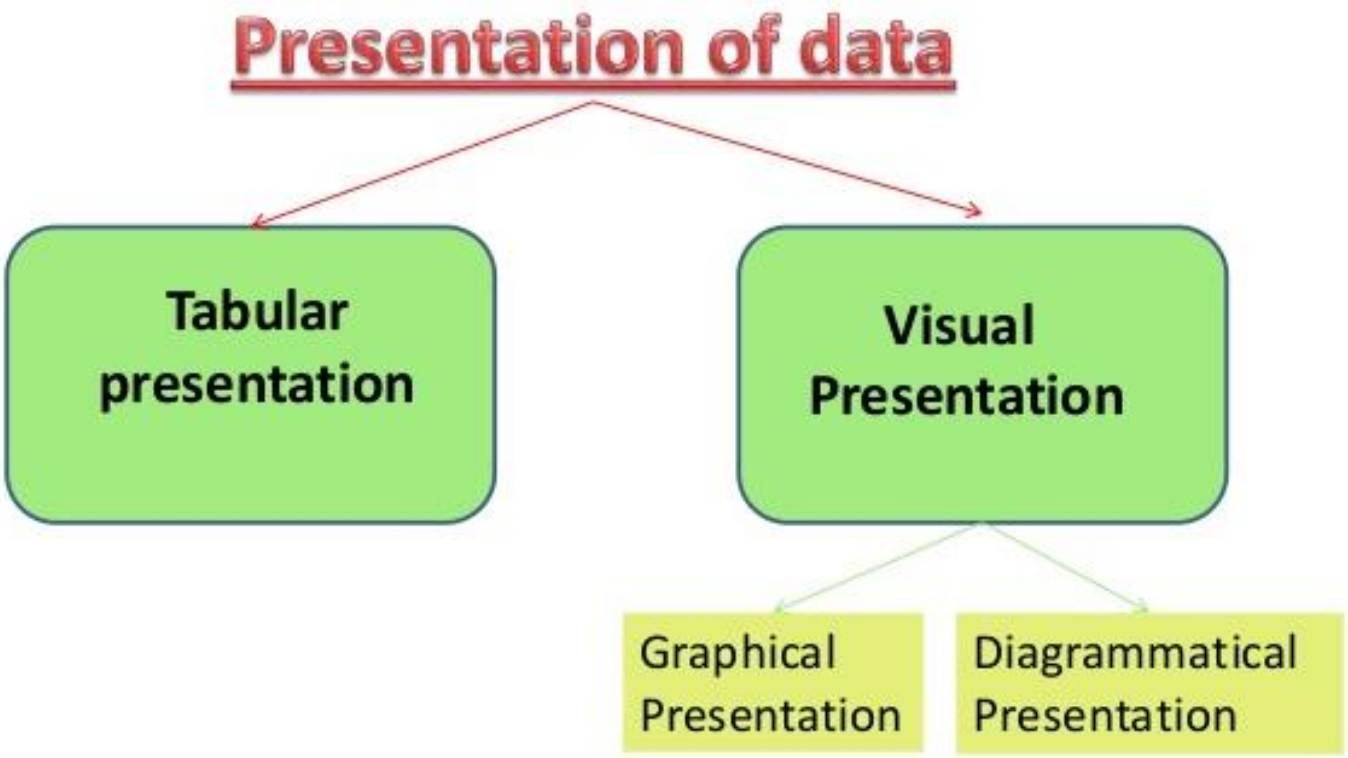
refers to the organization of **data** into tables, graphs or charts, so that logical and statistical conclusions can be derived from the collected measurements.

Data may be presented in(3 Methods):
-Textual
- Tabular or
- Graphical.

Text, tables, and graphs are effective communication media that **present** and convey **data** and information

Research Methodology

Scientific Publishing- Data representation



Research Methodology

Scientific Publishing- Data representation

A **Table** refers to any data which is presented in orderly rows across and/or down the page, often enclosed within borders.

A Figure refers to any other form of **presentation** such as a bar or pie chart, a graph, a diagram, a map, a photograph, a line drawing or a sample of material.

Research Methodology

Scientific Publishing- Data representation

Tabular **Presentation of data** is a method of **presentation of data**.

It is a systematic and logical arrangement of **data** in the form of Rows and Columns with respect to the characteristics of **data**.

It is an orderly arrangement which is compact and self-explanatory.

Research Methodology

Scientific Publishing- Data representation

•In a tabular **presentation**, **data** is arranged in columns and rows, and the positioning of **data** makes comprehension and understanding of **data** more accessible.

Table number. It is included for identification and becomes easy for reference in future.

•Title.

•Stub.

•Caption.

•Body.

•Footnote.

Research Methodology

Scientific Publishing- Data representation

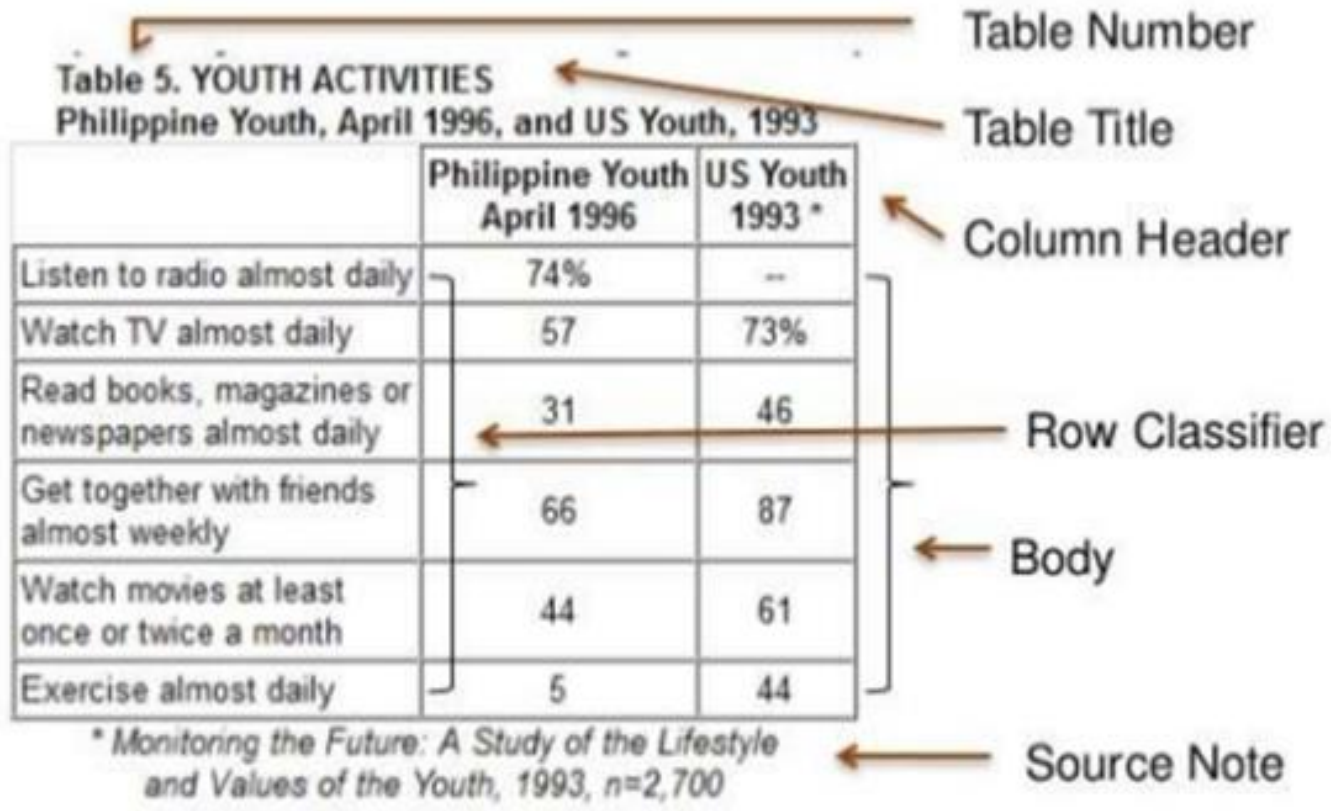
Table Number: .....
Title: .....
(Head Note, if any)

Stub (Row Heading)	Caption (Column Heading)				Total (Rows)
	Sub-head		Sub-head		
	Column-head	Column-head	Column-head	Column-head	
Stub Entries (Row Entries)		Body			
Total Columns					

Source Note:
Footnote:

Tabular Presentation of Data

Below is a sample of a table with all of its parts indicated:

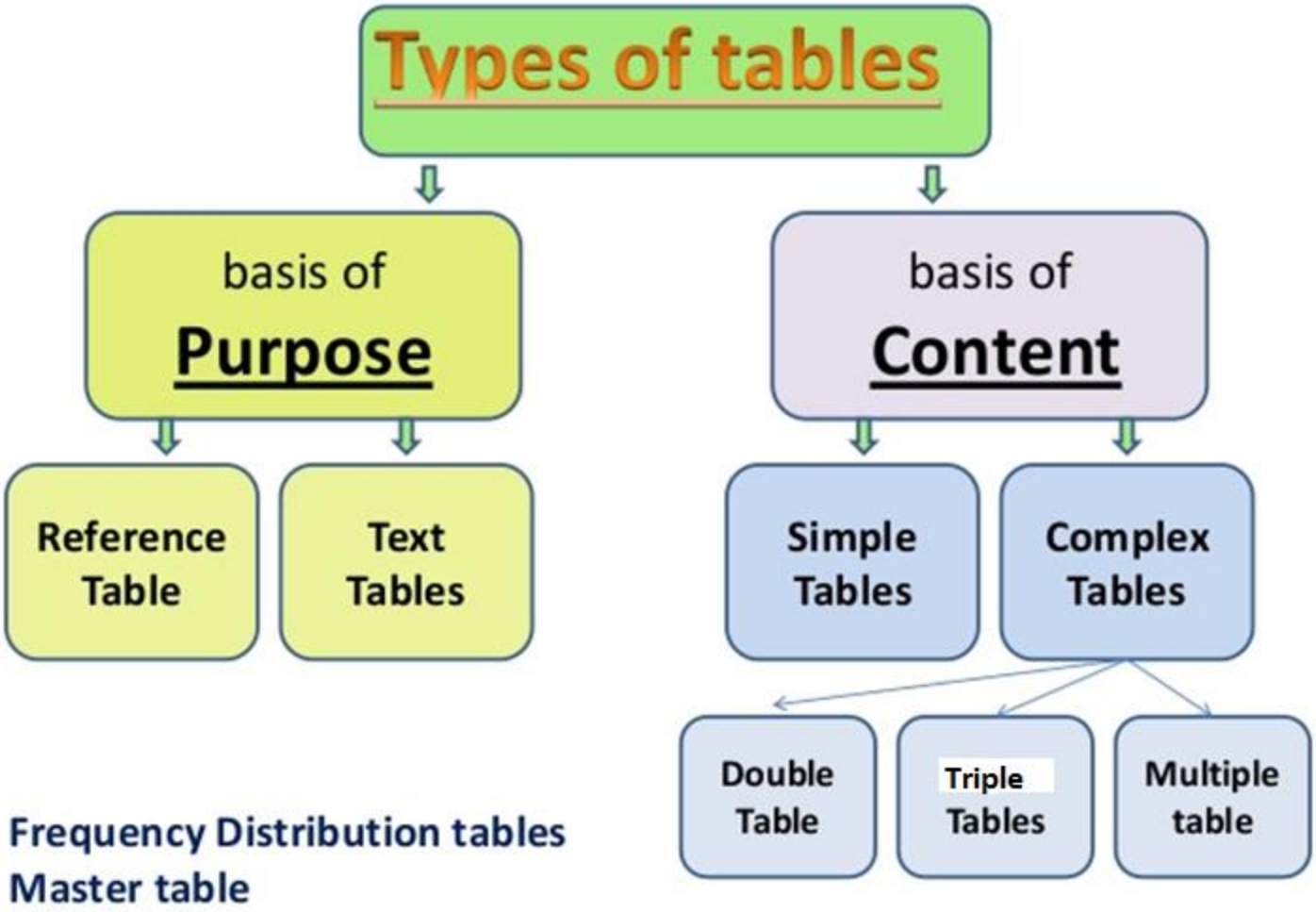


	Philippine Youth April 1996	US Youth 1993 *
Listen to radio almost daily	74%	--
Watch TV almost daily	57	73%
Read books, magazines or newspapers almost daily	31	46
Get together with friends almost weekly	66	87
Watch movies at least once or twice a month	44	61
Exercise almost daily	5	44

** Monitoring the Future: A Study of the Lifestyle and Values of the Youth, 1993, n=2,700*

Research Methodology

Scientific Publishing- Data representation



Advantages of table:

A **table** facilitates representation of even large amounts of **data** in an attractive, easy to read and organized manner.

The **data** is organized in rows and columns.

Table is one of the most widely used forms
of **presentation** of **data** since **data tables** are easy to construct and read.

One of the major **benefits of using** an Excel **table** is that it will automatically expand when you add a new record – even if it is added at the end of the **table**. So the range of cells that your name refers to will also automatically expand. This is known as a dynamic range.

Table and **its Characteristics**:

1. A **table** is perceived as a two-dimensional structure composed of rows and columns.

2. **Each table** row represents a single entity occurrence within the entity set.

3. **Each table** column represents an attribute, and **each** column has a distinct name.

Numerical Tables:

These are the most common **types of data**, which typically represent quantitative **data**, but sometimes may **present** a combination of quantitative and qualitative **data**.

As its name suggests, most of the body of the **table** consists of specific number values.

Features for good table

Attractive: It should be attractive as to leave **good** impression on reader.

Clarity: A **table** should be simple and clear i.e. can easily be understood.

Manageable size: Too much details should not be there and the size of the **table** should be medium i.e. neither too big nor too small.

Research Methodology

Scientific Publishing- Data representation

Before

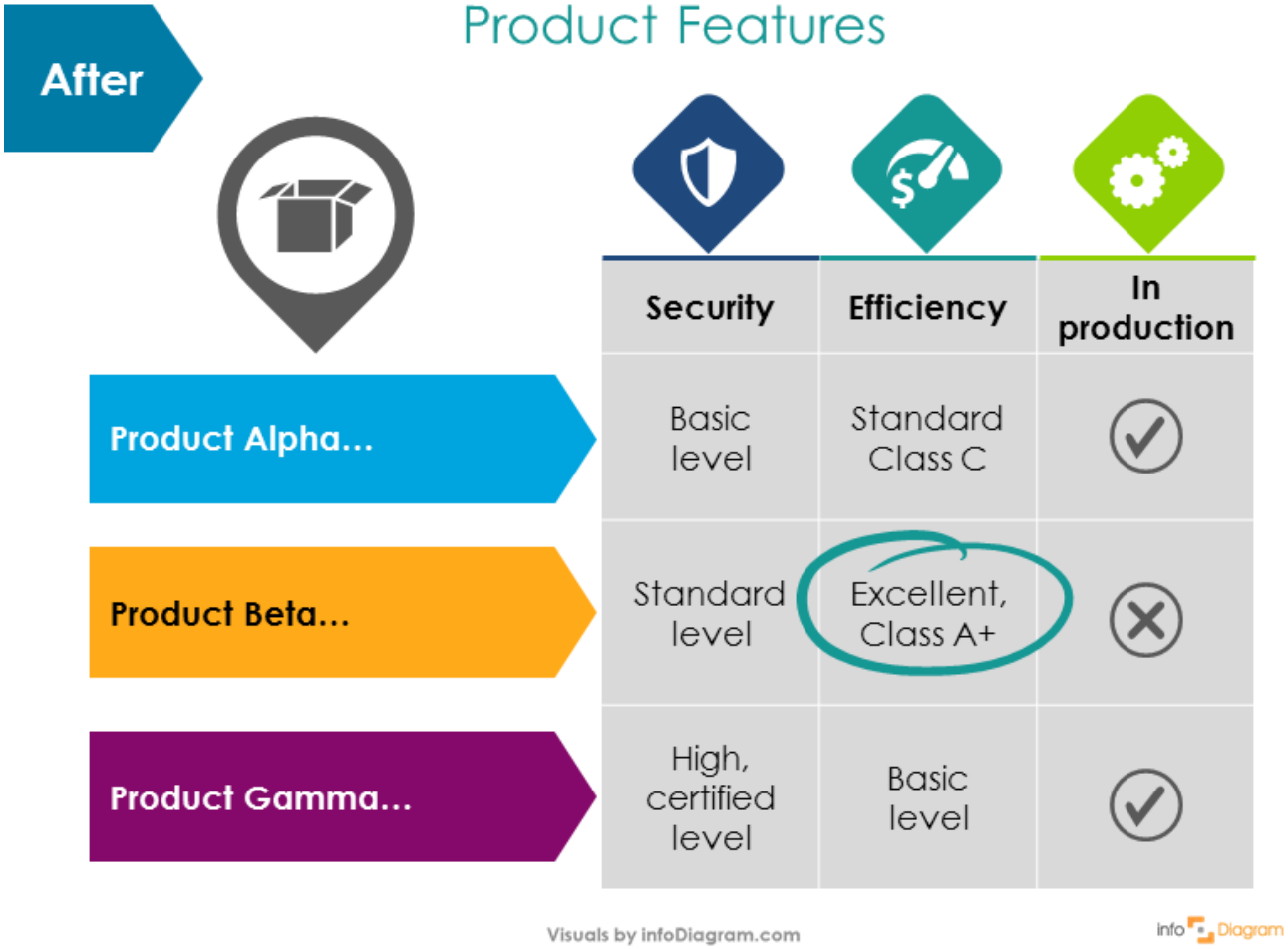
Product Features

	Security	Efficiency	In production
Product Alpha…	Basic level	Standard Class C	Yes
Product Beta…	Standard level	**Excellent, Class A+**	No
Product Gamma…	High, certified level	Basic level	Yes

Visuals by infoDiagram.com

Research Methodology

Scientific Publishing- Data representation



Research Methodology

Scientific Publishing- Data representation

Eg: Tables in census record, Appendices of Publications

Sl.No	Contents	Page numbers

Specific Heats of Common Materials

MATERIAL	SPECIFIC HEAT (Joules/gram • °C)
Liquid water	4.18
Solid water (ice)	2.11
Water vapor	2.00
Dry air	1.01
Basalt	0.84
Granite	0.79
Iron	0.45
Copper	0.38
Lead	0.13

Simple tables –

Data relating to only one characteristics

Gender	No of students
Boys	9
Girls	29

Double table -

Data relating to only 2 characteristics

Gender	Food habit	
	Vegetarians	Non Vegetarians
Boys	2	7
Girls	5	24

Triple table:

Data relating to only 3 characteristics

Gender	Food habit			
	Vegetarians		Non Vegetarians	
	Age below 20 years	Age 20 & above years	Age below 20 years	Age 20 & above years
Boys	0	2	1	6
Girls	1	4	10	14

Multiple table:

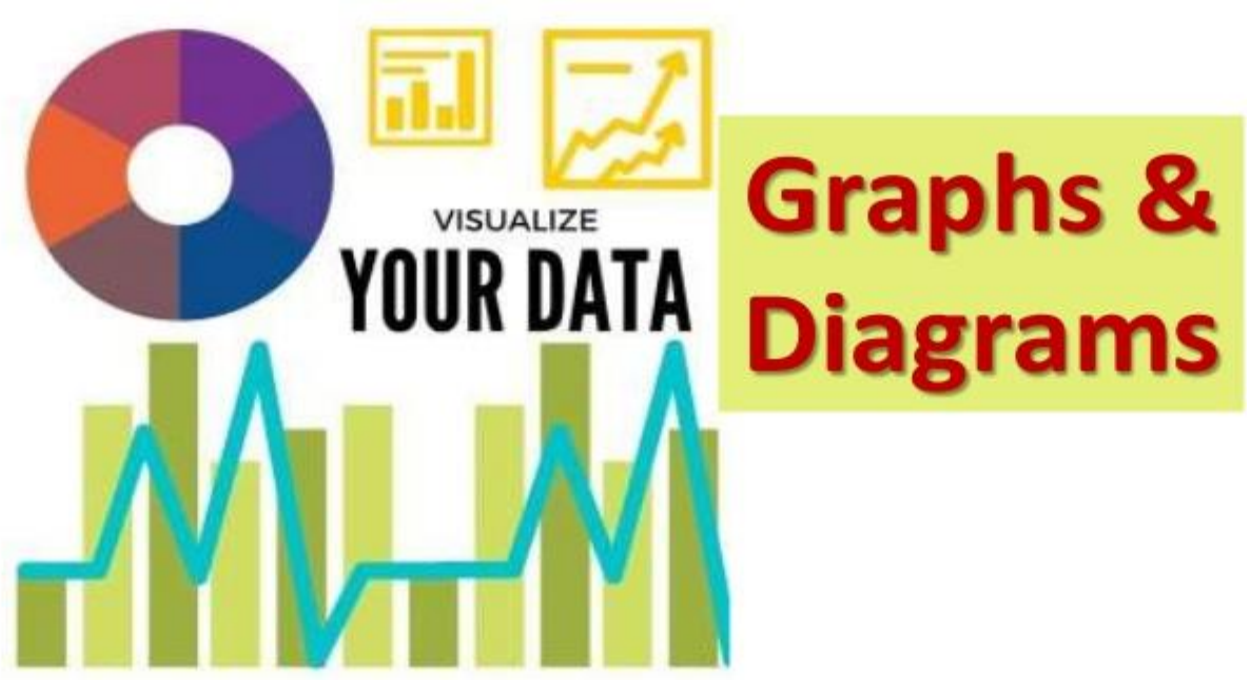
Gender		Food habit			
		Vegetarians		Non Vegetarians	
		Age <20 years	Age >=20 years	Age < 20 years	Age>= 20 years
Boys	Day scholars	0	0	1	4
	Hostellers	0	2	0	2
Girls	Day scholars	0	1	2	2
	Hostellers	1	3	8	12

Residing Area

← Age

Research Methodology

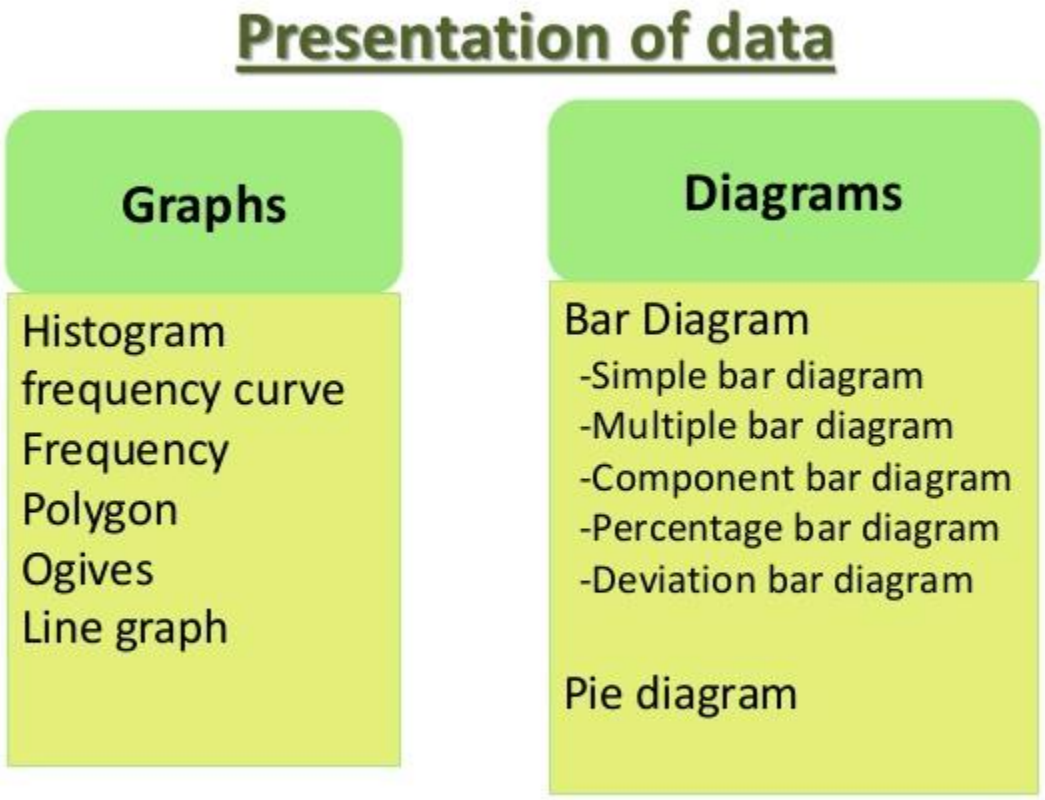
Scientific Publishing- Data representation



Research Methodology

Scientific Publishing- Data representation

Visualization, as the word suggests is the art of representing information in visual form like diagrams, charts or images. The visuals are usually supported by narration from the presenter.



RULES FOR DRAWING GRAPHS AND DIAGRAMS:

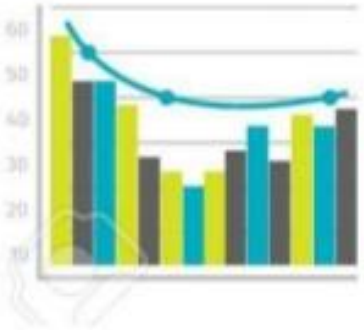
- First choose the form of diagrams /graphs which is capable of representing the given set of data.
- Title- gives information of diagrams or graphs contain.
- Scale – selection of scale should be neither too small or too large. The scale should also specify the size of unit and what it represents. (eg: No. of persons in thousands).
- Neatness
- Attractive – different types of lines or shades, colours etc can be used to make the pictures more attractive.
- Originality – helps the observer to see the details with accuracy
- Simplicity –good diagram depends upon ease with which the observer can interpret it.
- Economy – cost and labour should be exercised drawing a diagram.

Difference between Graphs and Diagrams:

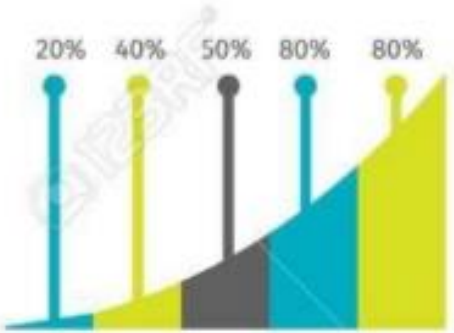
- To construct a graph, graph paper is generally used whereas a diagram is constructed on a plain paper.
- A graph represents mathematical relationship between two variables where as a diagram does not.
- Graphs are more appropriate than diagrams to represent frequency distributions and time series. Diagrams are not at all used for representing frequency distributions.
- Diagrams are more attractive to the eyes and as such are better suited for publicity and propaganda.
- Diagrams do not add anything to the meaning of the data and hence they are not helpful in analysis of data.
- Graphs are very much used by the statisticians and the research workers in their analysis.

Research Methodology

Scientific Publishing- Data representation



Graphs

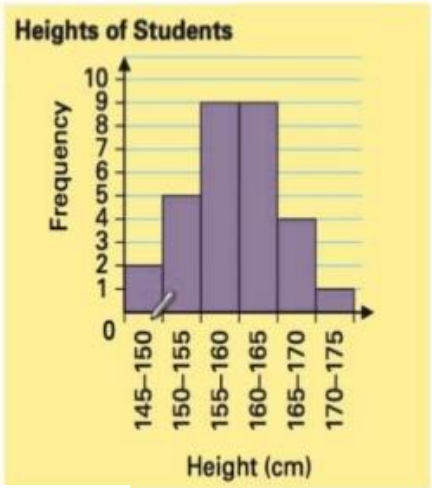


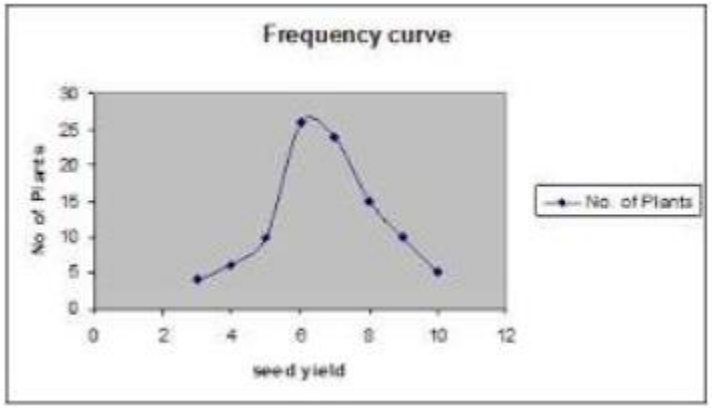
Research Methodology

Scientific Publishing- Data representation

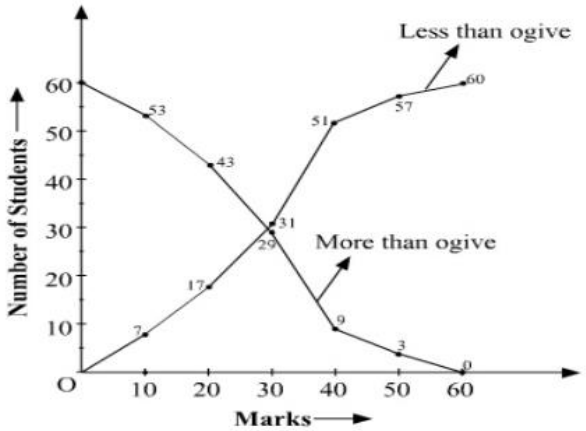
Histogram

Height (cm)	Frequency
145–150	2
150–155	5
155–160	9
160–165	9
165–170	4
170–175	1





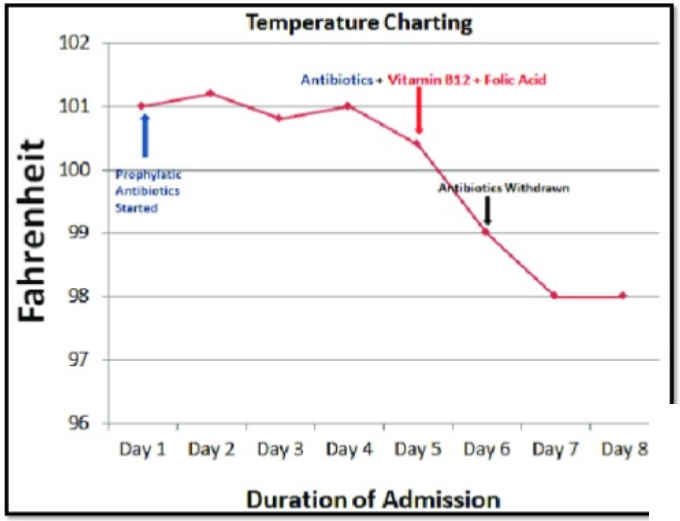
Ogives: (Cumulative Frequency Curves):



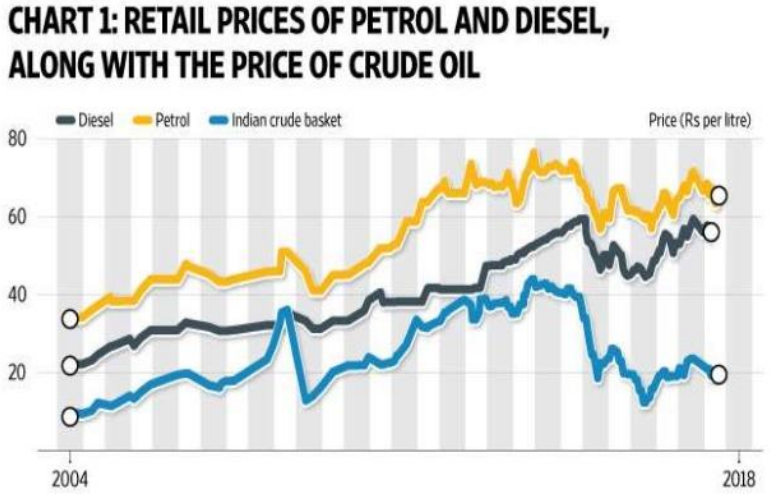
Research Methodology

Scientific Publishing- Data representation

Line Graph: (Time series graph)

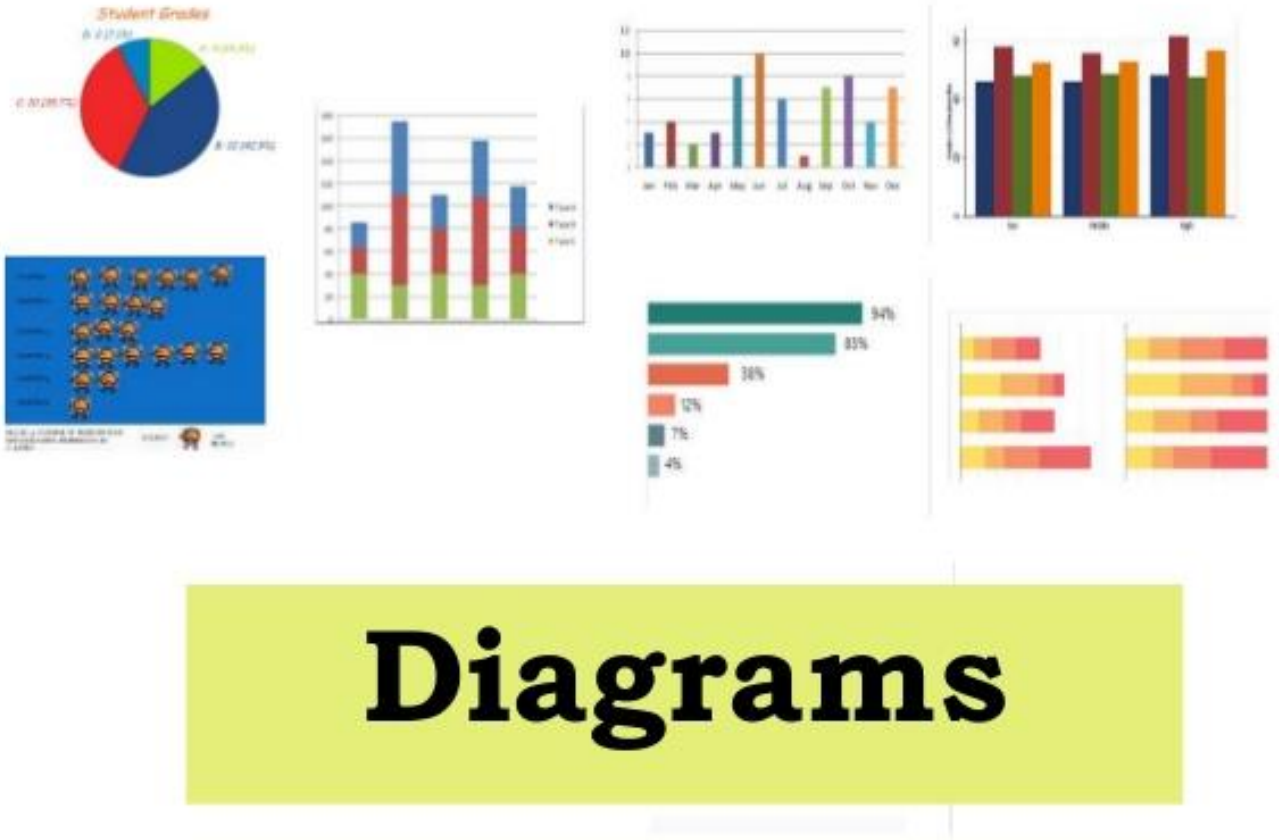


Line Graph: (Time series graph)



Research Methodology

Scientific Publishing- Data representation

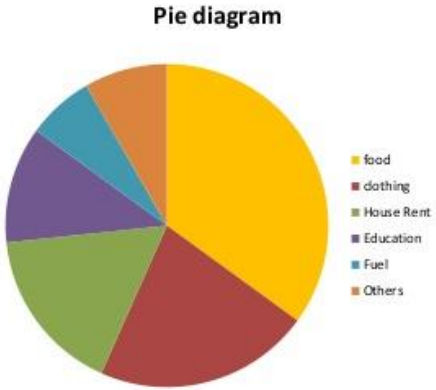


Research Methodology

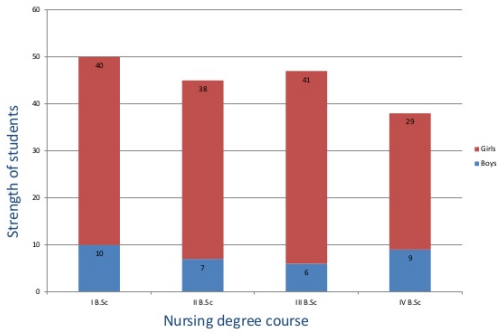
Scientific Publishing- Data representation

Eg: The following table gives the monthly expenditure of a family. It can be represented by means of a pie diagram.

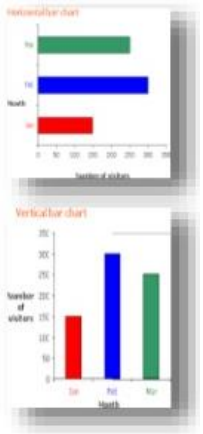
Items	Expenditure (Rs)	Degree measurement
Food	1050	126^0
Clothing	650	78^0
House rent	500	60^0
Education	350	42^0
Fuel	200	24^0
Others	250	30^0

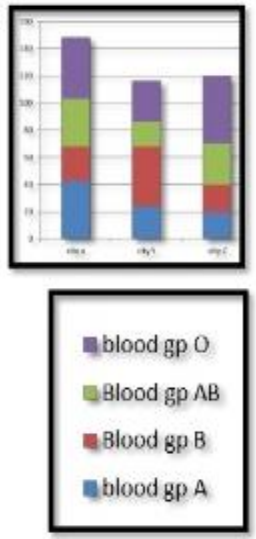


BAR DIAGRAM/ BARCHART



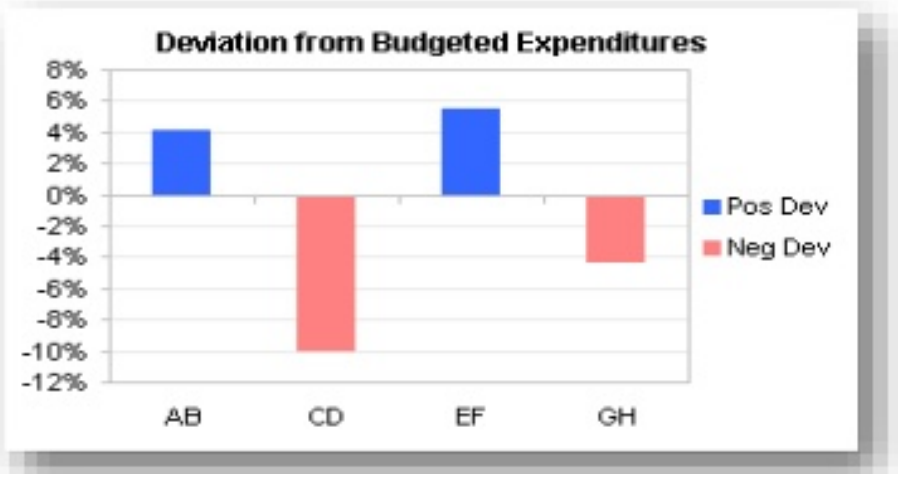
An example shows the strength of students in nursing degree course





Research Methodology

Scientific Publishing- Data representation



Research Methodology

Scientific Publishing- Data representation

ADVANTAGES:

- ✓ They are attractive
- ✓ They give a bird's eye-view of the data
- ✓ They can be easily understood by common men
- ✓ They facilitate comparison of various characteristics
- ✓ The impression created by them are long lasting
- ✓ Theorems and results of statistics can be visualized using graphs

Limitations:

- ✓They are visual aids. They cannot be considered as alternatives for numerical data.
- ✓Though theories and results could be easily visualized by diagrams and graphs, mathematical rigour cannot be brought in
- ✓Diagrams and graphs are not accurate as tabular data. Only tabular data can be used for further analysis.
- ✓By diagrammatical and graphical misrepresentation observers can be misled easily. It is possible to create wrong impressions using diagrams and graphs.

THANK YOU

RESEARCH METHODOLOGY

Topic: Results and Discussion (Data Interpretation)

Department of Computer Science & Engineering

DISCUSSION

Step by step:
An effective DISCUSSION

Results - Findings

- It describes what you found in your research, without **discussion**, **interpretation** or **reference to the literature**.

- Just the **facts**, presented as tables, figures, interview summaries and/or descriptions of what you found that is **important** and noteworthy.

- The objective is to present a **simple, clear and complete** account of the results of your research.

Discussion: is considered as the **heart** of the paper

Purpose: To state your

- Interpretations ;
- Opinions;
- Explain the implications of your findings and
- Make suggestions for future research.

Function:

- To answer questions posed in the Introduction,
- Explain how the results support the answers and
- How the answers fit in with existing knowledge on the topic.

Discussion

Not mere details about the results;
interpret and explain the results.

1. **(Un)expected results**
2. **Reference to previous research**
3. **Explanation**
4. **Exemplification**
5. **Deduction and hypothesis**
6. **Recommendation**

Provide
a **commentary** and not a reiteration of the results

Discussion

- Begin by briefly **summarizing** the previous chapters, then discuss what you found.
- Provide meaningful **answers** to the question
- Interpret **objectively** and **subjectively** and to **make references** to what others have said on the subject.
- Make sure that every **conclusion** you draw is **defensible** and not just your own personal opinion.

Discussion- Technique

1.

Organize the Discussion from the
specific to the general:
your findings to the literature, to theory, to practice

Discussion- Technique

2.

Begin by **re-stating the hypothesis** you were testing and answering the questions posed in the introduction

Discussion- Technique

3.

- Explain how your results relate to expectations
- Clearly state why they are acceptable and
- How they are consistent or fit with published knowledge

Discussion- Technique

4.

Address **all** the results regardless of whether or not the findings were statistically significant.

Discussion- Technique

5.

Describe the patterns, principles, and relationships shown by each major finding/result and put them in perspective.

The sequencing:

First - state the answer,

Second - support with relevant results,

Third - cite the work of others.

Discussion- Technique

6.

Defend your answers by explaining both why your answer is satisfactory and why others are not.

Only by giving both sides to the argument can you make your explanation convincing.

Discussion- Technique

7.

Discuss and evaluate conflicting explanations of the results.

This is the sign of a good discussion.

Discussion- Technique

8.

Discuss any unexpected findings.

When discussing an unexpected finding, begin the paragraph with the finding and then describe it.

Discussion- Technique

9.

Identify potential **limitations** and weaknesses and comment on the relative importance of these to your interpretation of the results and how they may affect the validity of the findings.

When identifying limitations and weaknesses, avoid using an apologetic tone.

Discussion- Technique

10.

- Summarize concisely (brief, and specific)
- Explain implication and importance
- Provide recommendations (not >2) for further research.

Discussion- Do’s & Don’ts

DO: Provide context and explain why people should care.

DON'T: Simply rehash your results.

DO: Emphasize the positive.

DON'T: Exaggerate.

DO: Look toward the future.

DON'T: End with it.

RESEARCH METHODOLOGY

Topic: Summary and Conclusions

Department of Computer Science & Engineering

Summary, Conclusions & Recommendation

Summary, Conclusions & Recommendation

- **The summary is a brief restatement of the main findings presented under each factor**

- **The conclusion is an interpretation of the facts you gathered and discussed.**

  It is not a repetition of the facts.

  It is not an action that one must take.

Summary

- The Summary section may be the Conclusion
- Summary: summarizes the findings/conclusion
- Conclusion: ultimate take-away message
- Future work
- Limitation

The Purpose of Conclusion

1. Tie together, integrate and synthesize the various issues raised in the discussion sections, while reflecting the -Introduction, Problem Statement or Objectives

2. Provide answers to the research question (s)

3. Identify the theoretical implications of the study

4. Highlights the study limitations

5. Provide direction and areas for future research

Conclusion

Succinctly summarize implications

No sweeping statements or conclusions that reach beyond your data

Present the bottom line message, point, value of the described study

Tell the reader what they should take away

- **Advantages**
- **Novelty**
- **Limitations**
- **Suggestions**

The content of a good conclusion

- Be a logical ending synthesizing what has been previously discussed and never contain any new information or material

- It must pull together all of the parts of your argument and refer the reader back to the focus you have outlined in your introduction and to the central topic and thereby create a sense of unity.

- Be very systematic, brief and never contain any new information

- Add to the overall quality and impact of the research.

The content of a good conclusion

Restate the research questions which reinforces the importance of the study and its findings.

Empirical Findings: summary of the main finding in the different chapters provide answers to or the specific research

Theoretical Implication: **P**resent a modest position of how the work has contributed to existing understanding of concepts that has been investigated.

Recommendation for future research: **F**urther research that has not been covered but is worthwhile to investigate in the near future.

Limitation of the study: **I**dentify the various limitations which were encountered during the sampling, lab work, data collection and analysis stages of the research or project.

Different Styles Of Referencing

Agenda .....

- **Objective.**
- **What is reference style.**
- **Why to reference.**
- **Types of references.**
- **Different styles of writing reference.**

  A. Harvard style of referencing.

  B. American Psychological Association style (APA) .

  C. Vancouver style.

  D. MLA citation style (modern language association).

  E. The Chicago manual of style .

  F. Royal society of chemistry style

- **Conclusion**

style ????

- A referencing style is a specific format for presenting in-text references (footnotes or endnotes),

  and bibliography.
- It is a act of referring.

Reference :

- The action of mentioning or alluding to something or,
- The use of a source of information in order to ascertain something.

Why to reference??

- Proves that substantial research has been done to support our analysis .
- Enables others to follow up on our work .
- Gives credit to other people's work .
- Avoids charges of plagiarism.
- Required to support all significant statements.
- Used to indicate the origin of material & source for research & further reading.

Types of references

- Journal Reference
- Book Reference
- Internet Reference

Reference Elements

- Authors name
- Article title
- Journal name
- Year
- Volume
- Page numbers

Different styles of writing references:

- Harvard style of referencing.
- American Psychological Association style (APA) .
- Vancouver style.
- MLA citation style (modern language association).
- The Chicago manual of style .
- Royal society of chemistry style.

Harvard style of referencing

- Author's name followed by its initials.
- Year of publication.
- Article title with single quotation mark followed by full stop.
- Name of Journal in italic form.
- Volume followed by a comma
- Issue no. in bracket.
- Page no.

➢Example

1. Padda, J. (2003) 'creative writing in coventry'. *Journal of writing studies 3* (2), 44-59.
2. Lennernas, H. (1995) 'Experimental estimation of the effective unstirred water layer thickness in the human jejunum & its importance in oral drug absorption'. *Eur. J. pharm sci* (3), 247-253.

Vancouver style.

- Author Surname followed by Initials.
- Title of article followed by double quotation.
- Title of journal (abbreviated).
- Date of Publication followed by semicolon.
- Volume Number.
- Issue Number in bracket.
- Page Number.

➢ Example

1. Haas AN, Susin C, Albandar JM, et al. "Azithromycin as a adjunctive treatment of aggressive periodontitis: 12-months randomized clinical trial". N Engl J Med. 2008 Aug; 35(8):696-704.

✓ Vancouver Style does not use the full journal name, only the commonly-used abbreviation: “New England Journal of Medicine” is cited as “N Engl J Med”.

MLA citation style (modern language association)

- Authors name.
- Title of article.
- Name of journal.
- Volume number followed by decimal & issue no.
- Year of publication.
- Page numbers.
- Medium of publication.

➢ Example

1. Matarrita-Cascante, David. "Beyond Growth: Reaching Tourism-Led Development." *Annals of Tourism Research* 37.4 (2010): 1141-63. Print

American Psychological Association style

- Author's name followed by its initials.
- Year of publication.
- Article title followed by full stop.
- Name of Journal in italic form
- Volume followed by a comma
- Page no.

➢ Example

1. Alibali, M. W., Phillips, K. M., & Fischer, A. D. (2009). Learning new problem-solving strategies leads to changes in problem representation. *Cognitive Development,* 24, 89-101.

The Chicago manual of style

- Name of author.
- Article title in double quotation mark.
- Title of journal in italic.
- Volume.
- Year of publication.
- Page no.

➢Example

1. Joshua I. Weinstein, “The Market in Plato’s ” *Classical Philology,* 104 (2009): 440.

Royal society of chemistry styling

- INITIALS. Author's surname.
- Title of journal (abbreviated).
- Year of publication.
- Volume number.
- Pages no.

➢ Example

H. Yano, K. Abe, M. Nogi, A. N. Nakagaito, *J. Mater. Sci*., 2010, 45, 1–33.

Difference between Reference List and Bibliography

- **Reference list** – sources we have cited in our text arranged in the order they appeared within the text. It is usually put at the end of our work but it can also appear as a footnote (at the bottom of the page), or endnote (at the end of each chapter) which serves a similar purpose.

- **Bibliography** – a separate list of sources we have consulted but not specifically cited in our work including background reading. It is arranged alphabetically by the author's surname.

Conclusion

- We conclude that there are many standard style used for referencing, we can use any one of them.
- It gives us a standard format of presenting or reference.
- Supports or significant statement and helps to know origin of work.
- Plagiarism can be avoided.

Reference

- Art Of Writing & Publishing In Pharmaceutical Journals By Ajay Semalty, Shaiiendra K. Saraf, Mona Semalty, Shubhini A. Saraf, Ranjit Singh, 1st Edition: Pharma Book Syndicate, Hyderabad, Pg. No. 80.
- Library Services Help Sheet, London South Bank University, Perry Library & Learning Resources Pg. No. 2.
- Different Style Of Writing References In A Research Report By Caryn Anderson.
- Coventry University Harvard Reference Style Guide By Lisa Ganobcsik Williams & Catalina Neculai, Pg. No. 7.