RESEARCH METHODOLOGY

Unit-03:

Testing of Hypotheses and Data Analysis

Raghu B. A
Priya B.
Santhosh Kumar V
Department of Computer Science & Engineering

RESEARCH METHODOLOGY

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

Santhosh Kumar V
Department of Computer Science & Engineering

Introduction

- Principal Instrument of research
- Function is to suggest experiments and observation.

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

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.*"

Examples

"Students who receive counselling will show a greater increase in creativity than students not receiving counselling"

"The automobile A is performing as well as automobile B".

Table 1.1 The Effect of Aspirin on Heart Attacks

Condition	Heart Attack	No Heart Attack	Attacks per 1000
Aspirin	104	10,933	9.42
Placebo	189	10,845	17.13

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".*

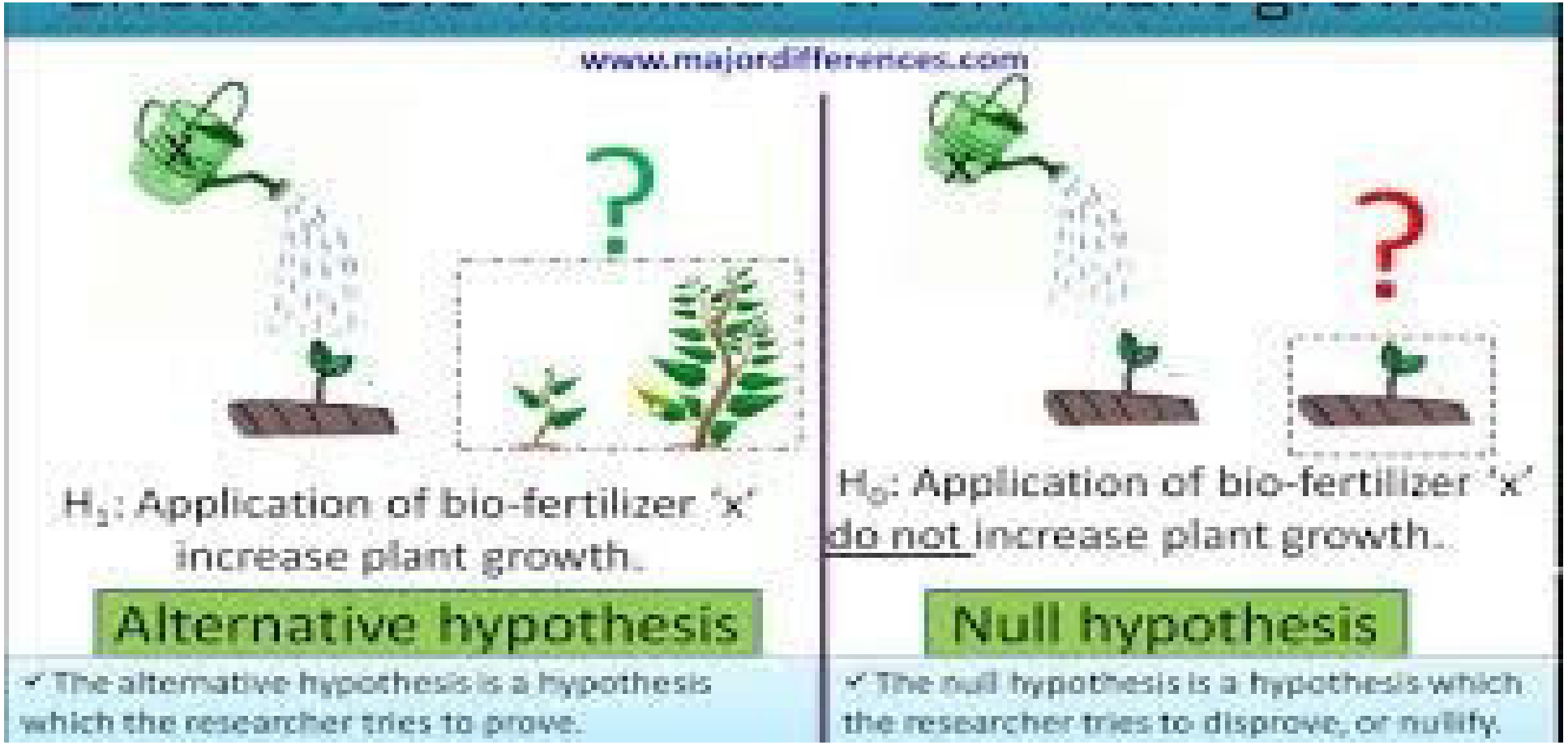
Alternate Hypothesis – If method A is superior than B (H_1).

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

Table 12.2 Data for Example 1 with Percentage and Rate Added

	Heart Attack	No Heart Attack	Total	Heart Attacks (%)	Rate per 1000
Aspirin	104	10,933	11,037	0.94	9.4
Placebo	189	10,845	11,034	1.71	17.1
Total	293	21,778	22,071		

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.

Example

Doctors recommend teenagers between 14-18 years to get at least 8 hrs sleep for proper health.

Authorities suspect that students at their school are getting less than 8 hours sleep on average.

To test this, we randomly take sample of 42 students and ask them how much sleep they get per night.

Mean = 7.5 hours.
Alternate H_1: avg amt of sleep student gets is < 8 hrs
$H_0 : \mu >= 8$

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:

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

Possible alternate hypothesis

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

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)

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

The 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.

Level of Significance

The significance level, also denoted as α, is the probability of rejecting the null hypothesis when it is true

Ex : a significance level of 0.05 indicates a 5% risk of concluding that a difference exists when there is no actual difference

Type 1 and Type 2 errors

Type 1 error

If Null hypothesis is rejected when it is *true*

Type 2 error.

If Null hypothesis is accepted when it is *not true*

In other words

Type1 means – rejection of hypothesis when should have been accepted and

Type 2 means accepting hypothesis when should have been rejected.

Type 1 and Type 2 Errors

What are these errors?

- These are errors that arise when performing hypothesis testing and decision making

- **Type 1 error** (*false positive* conclusion)
  - Stating difference when there is no difference, alpha
  - Related to p value, ***how?***
  - Set at 1/20 or 0.05 or 5%
  - The probability is distributed at the tails of the normal curve i.e., 0.025 on either tail

- **Type 2 error** (*false negative* conclusion)
  - Stating no difference when there is a difference, beta
  - Occurs when sample size is too small.
  - Conventional values are 0.1 or 0.2
  - Related to power, ***how?***

Type 1 and Type 2 Errors

Example 1

Null Hypothesis	Type I Error / False Positive	Type II Error / False Negative
Person is not guilty of the crime	Person is judged as **guilty** when the person actually **did not** commit the crime (convicting an innocent person)	Person is judged **not guilty** when they actually **did** commit the crime (letting a guilty person go free)
Cost Assessment	Social costs of sending an innocent person to prison and denying them their personal freedoms (which in our society, is considered an almost unbearable cost)	Risks of letting a guilty criminal roam the streets and committing future crimes

Type 1 and Type 2 Errors

Example 2

Null Hypothesis	Type I Error / False Positive	Type II Error / False Negative
Wolf is not present	Shepherd thinks wolf is present (shepherd cries wolf) when no wolf is actually present	Shepherd thinks wolf is NOT present (shepherd does nothing) when a wolf is actually present
Cost Assessment	Costs (actual costs plus shepherd credibility) associated with scrambling the townsfolk to kill the non-existing wolf	Replacement cost for the sheep eaten by the wolf, and replacement cost for hiring a new shepherd

Type 1 and Type 2 Errors

Example 3

Null Hypothesis	Type I Error / False Positive	Type II Error / False Negative
Medicine A cures Disease B	(H_0 **true**, but rejected as **false**)Medicine A **cures** Disease B, but is **rejected as false**	(H_0 **false**, but accepted as **true**)Medicine A **does not cure** Disease B, but is **accepted as true**
Cost Assessment	Lost opportunity cost for rejecting an effective drug that could cure Disease B	Unexpected side effects (maybe even death) for using a drug that is not effective

Type 1 and Type 2 Errors

Possible Errors in Hypothesis Test Decision Making *(continued)*

Possible Hypothesis Test Outcomes		
	Actual Situation	
Decision	H_0 True	H_0 False
Do Not Reject H_0	No Error Probability 1 - α	Type II Error Probability β
Reject H_0	Type I Error Probability α	No Error Probability 1 - β

Type 1 and Type 2 Errors

Possible Errors in Hypothesis Test Decision Making *(continued)*

Possible Hypothesis Test Outcomes		
	Actual Situation	
Decision	H_0 True	H_0 False
Do Not Reject H_0	No Error Probability 1 - α	Type II Error Probability β
Reject H_0	Type I Error Probability α	No Error Probability 1 - β

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

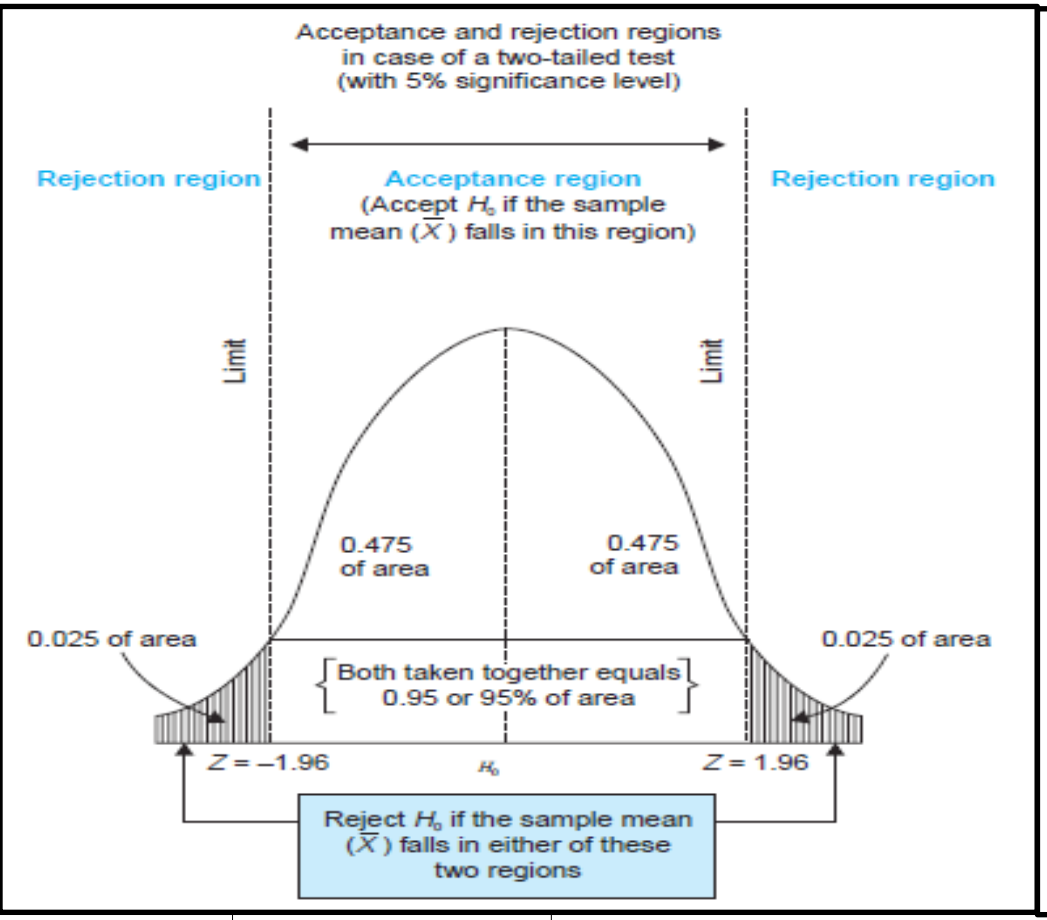
$H_0:\mu >= \mu_{H0}$ Aganist $H_a:\mu < \mu_{H0}$

If we have ≠ in alternate hypotheses – Two tailed test

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

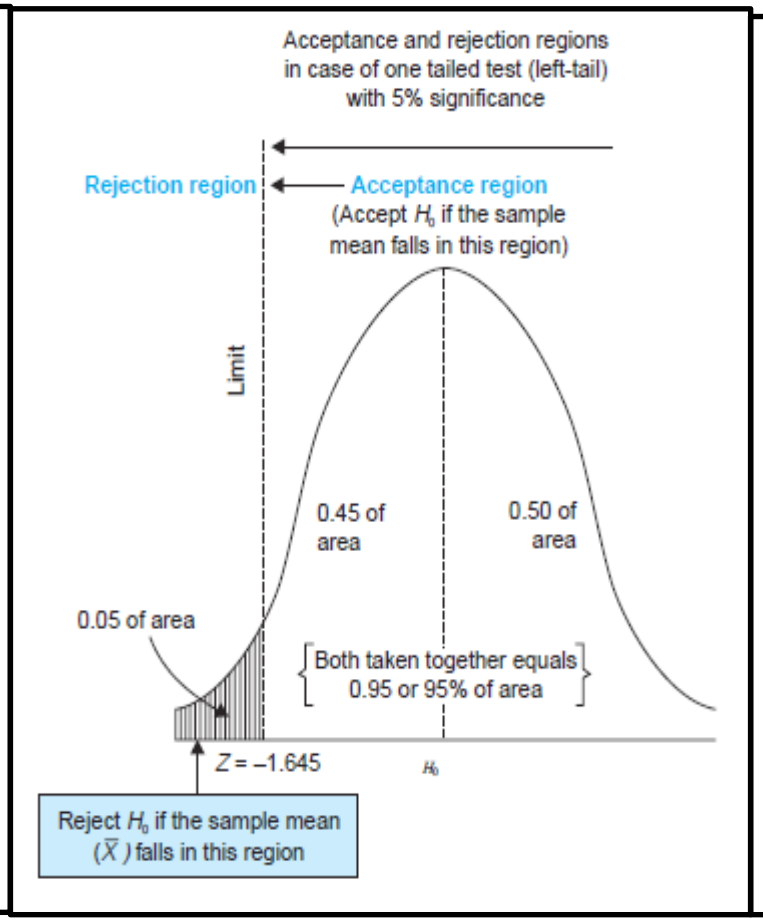
If we have < sign in alternate hypotheses – left **tailed**

One tailed and two tailed test



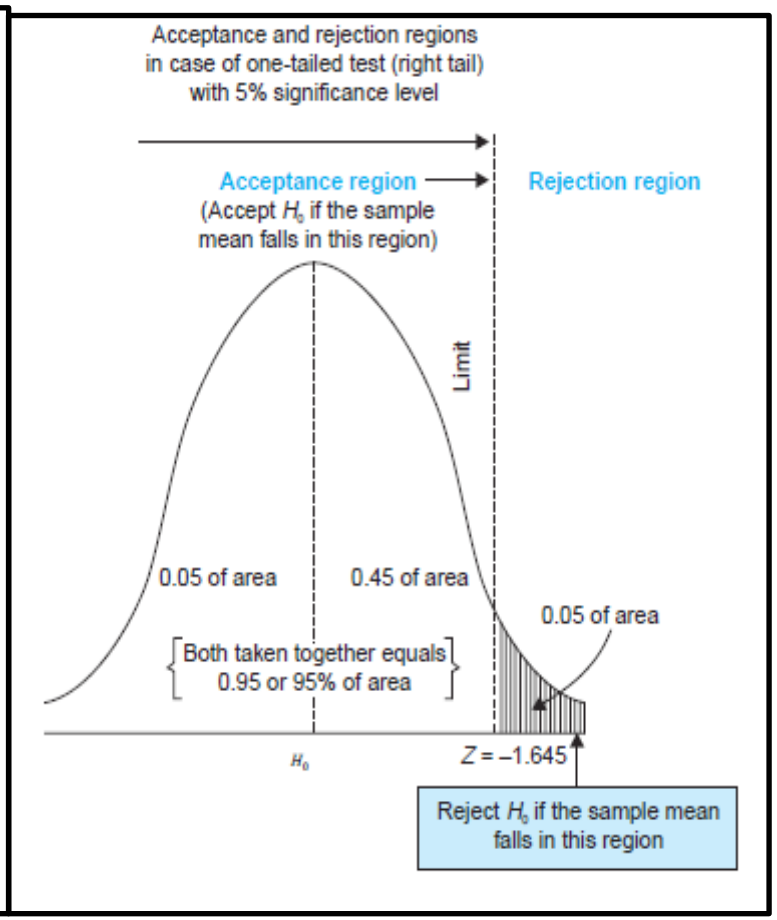
$H_0: \mu = \mu_{H0}$

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



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

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



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

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

Steps in Hypothesis Testing

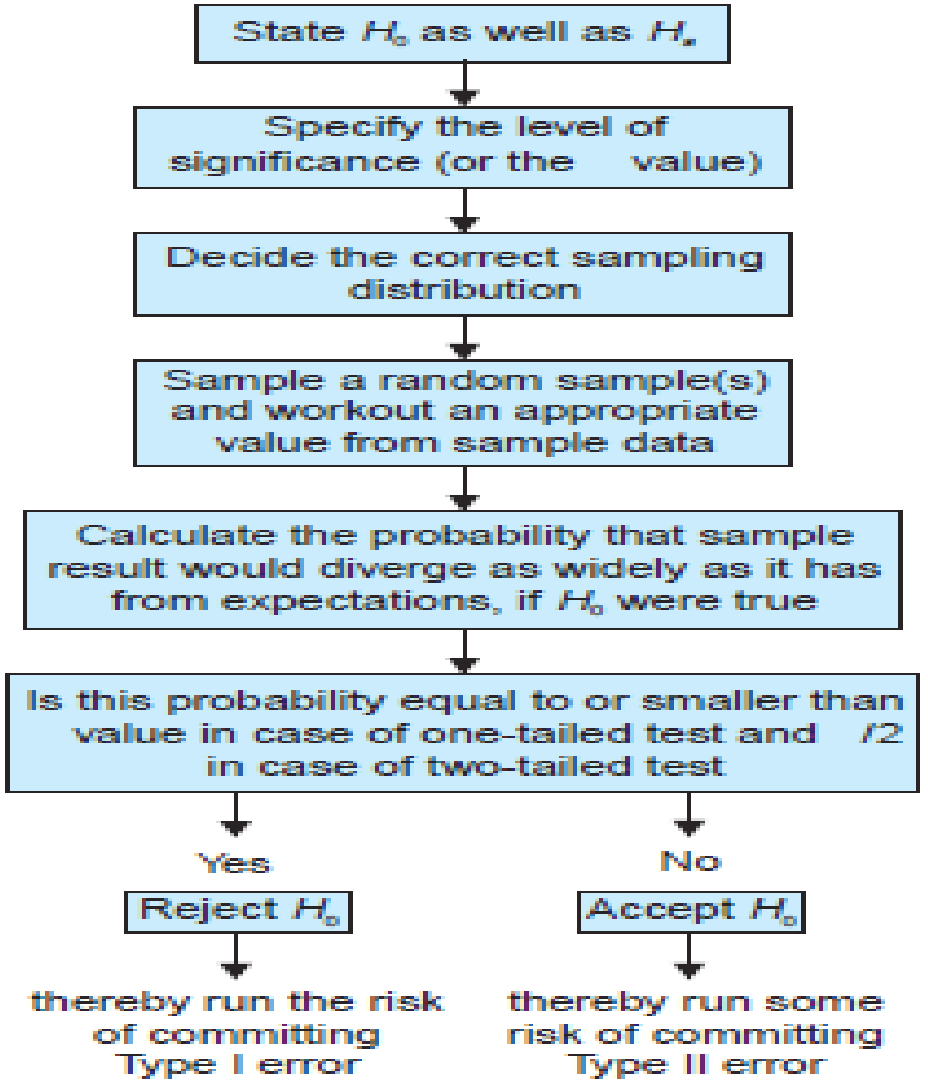
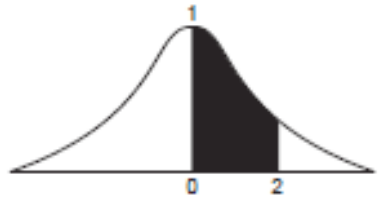


Table 1: Area Under Normal Curve

An entry in the table is the proportion under the entire curve which is between $z = 0$ and a positive value of z. Areas for negative values for z are obtained by symmetry.



Areas of a standard normal distribution

z	.0	0.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
.7	.2580	.2611	.2642	.2673	.2903	.2734	.2764	.2794	.2823	.2852
.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986

Table 1: Area Under Normal Curve

An entry in the table is the proportion under the entire curve which is between $z = 0$ and a positive value of z. Areas for negative values for z are obtained by symmetry.

z	.0	0.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
.7	.2580	.2611	.2642	.2673	.2903	.2734	.2764	.2794	.2823	.2852
.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441

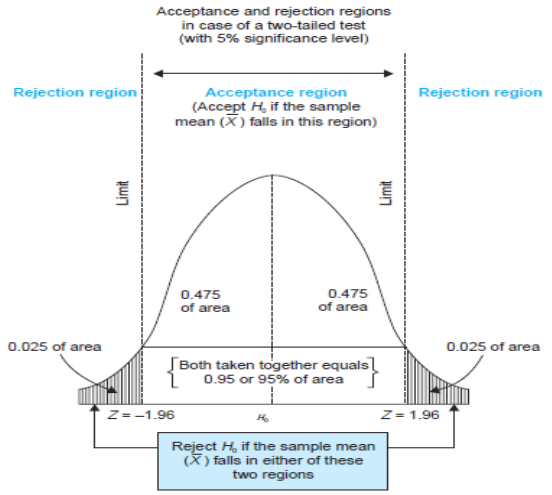
z	.0	0.01	.02	.03	.04	.05	.06	.07	.08	.09
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990

Eg. Hypothesis Testing

The average IQ for the adult population is 100 with a standard deviation of 15. A researcher believes that this value has changed. So a IQ test is conducted on 75 random adults, resulting in avg IQ of 105.

i) Is there enough evidence to suggest that the avg IQ has changed. (Assume α = 5%)

ii) What is the power of the test for μ = 105.



1. State H0 and HA — H0 = μ = μH0 = 100 Ha = μ ≠ 100
2. Specify α — α = 5%
3. Choose sampling distribution & critical value (based on α) — Z distribution : 2-tailed:
4. Calculate test statistic (Z)
5. Calculate Probability (P)
6. P < α (one tailed)
   P < α/2 (two tailed)
   Yes => Reject H_0
   (Statistically Significant)

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{105 - 100}{\frac{15}{\sqrt{75}}} = 2.89$$

Since > i.e 2.89 > 1.96
There is evidence to reject H0.
There is evidence that IQ has changed.

= 1 - 0= 1-[0.5+ = 0.4981 = 0.0091 < 0.025

No => Accept H_a

Since
There is evidence to reject H0.
There is evidence that IQ has changed.

Eg. Hypothesis Testing

The average IQ for the adult population is 100 with a standard deviation of 15. A researcher believes that this value has changed. So a IQ test is conducted on 75 random adults, resulting in avg IQ of 105.

i) Is there enough evidence to suggest that the avg IQ has changed. (Assume α = 5%)

ii) What is the power of the test for μ = 105.

1. State H0 and HA
2. Specify α
3. Choose sampling distribution
4. Calculate test statistic (Z_c)
5. Calculate Probability (P)
6. P < α (one tailed)
   P < α/2 (two tailed)
   Yes => Reject H_0
   (Statistically Significant)

   No => Accept H_a

Eg. Hypothesis Testing

A chemical process produces 15 lbs or less of waste for every 60lb batch, with a SD of 5 lbs. A random sample of 100 batch gave an average waste of 16 lbs per batch.

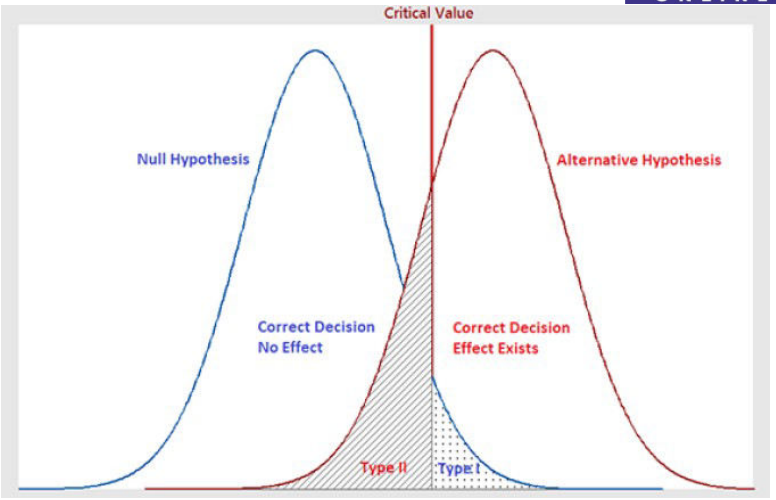
i) Has the wastage increased at a significance level of 10%.
ii) Compute the power of the test for μ = 16.
iii) If the significance level is increased to 20%, what is the new power of the test for μ = 16.?

1. State H0 and HA
2. Specify α
3. Choose sampling distribution
4. Calculate test statistic (Z_c)
5. Calculate Probability (P)
6. P < α (one tailed)
   P < α/2 (two tailed)
   Yes => Reject H_0
   (Statistically Significant)

   No => Accept H_a

Statistical Power of Hypothesis Test

H0 : no effect/no change
Ha : effect/change



Type 1 Error (α) = Prob(Reject H0| H0 is True)
Type 2 Error (β) = P (not Rejecting H0 | H0 is False)

Accepting Null Hypothesis when its should be Rejected.
Failure to choose Ha when Ha is True.
There is "no effect" when in reality the is "effect"
Hypothesis Test is not able to "detect a change", where as in reality there is a "change"
False Negative: Test result says "No evidence to reject H0" (Accept H0)

Eg: Hypothesis test says: Medicine is "not effective" when its actually effective.

β = Failure to choose Ha when Ha is True. (desirable to be a low value)

Power of Hypothesis Test (1- β) : The **power** of a test is the probability of making the correct decision when the alternative hypothesis is true.
Power is the ability of the test to detect an effect that exists in the population.

High Power is desirable (>= 80%)

Statistical Power of Hypothesis Test

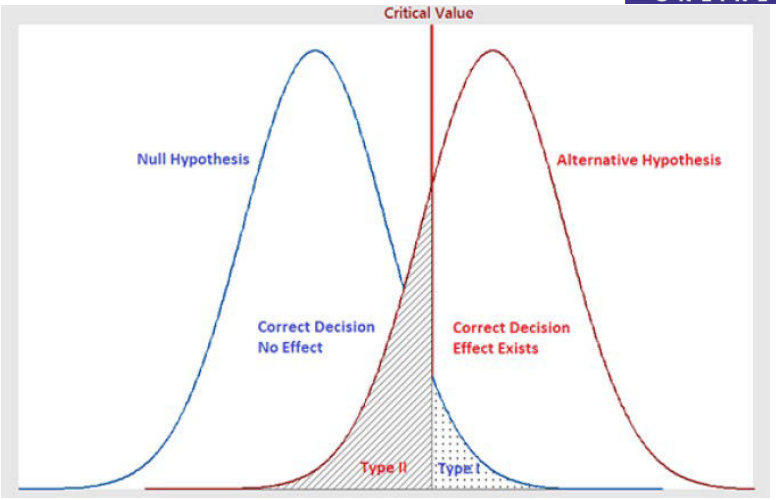
H0 : no effect/no change
Ha : effect/change

β = Failure to choose Ha when Ha is True. (desirable to be a low value)

Power of Hypothesis Test (1- β) : The **power** of a test is the probability of making the correct decision when the alternative hypothesis is true.

Power is the ability(likelihood) of the test to detect an effect that exists in the population.

High Power is desirable (>= 80%)



Procedure:

Do Hypothesis test at a significance level (α) (eg. =5%, 1%)
Calculate the Power (1- β) of the test.
if its acceptable (>= 80%), then sample size is ok.
Otherwise increase sample size

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:

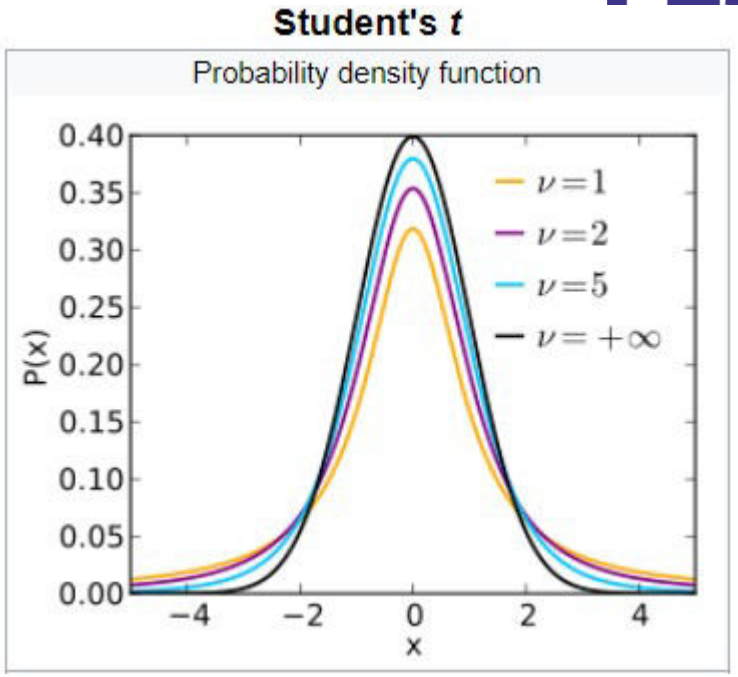
$$z = \frac{\bar{X} - \mu_{H_0}}{\sigma_p / \sqrt{n}}$$

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:

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

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



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).

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

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

Chi-Square tests

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.

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

Conditions for chi-square test.

- Observations must be random and independent
- No group should have freq < 10. When freq are less than 10, group the adjoining groups
- Overall no must be large (> 50)
- Constrains must be linear

Degree of Freedom

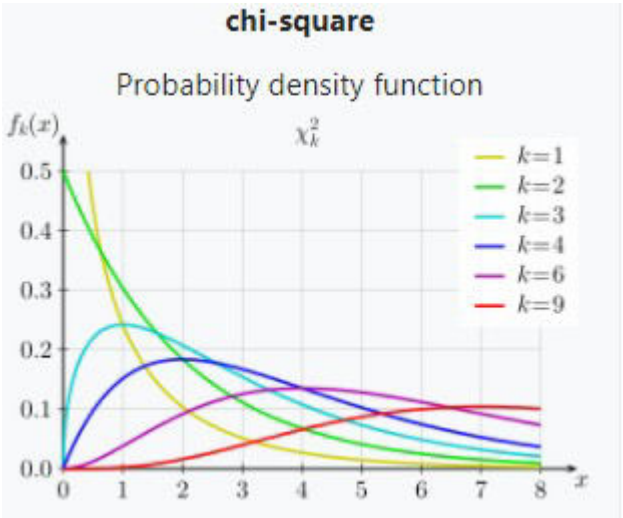
Number of Independent value which are assigned to statistical distribution (n-1) . Eg: Tossing of a die 132 times

λ

Num on Top	1	2	3	4	5	6	Total
Observed Frequency	16	20	25	14	29	28	132

Number of Independent value which are assigned to statistical distribution ((r-1)(c-1)

Observed Frequency	Party A	Party B	Row Total
Male	55	65	120 (M)
Female	50	30	80
Col Total	105 (A)	95	200 (N)



Observed Frequency vs Expected Frequency

=Observed frequency in i^{th} row and j^{th} column

=Expected frequency in i^{th} row and j^{th} column.

=

Observed Frequency	Party A	Party B	Row Total
Male	55	65	120 (M)
Female	50	30	80
Col Total	105 (A)	95	200 (N)

$P(M) = ¿$

$P(A) = ¿$

$P(A \cap B) = ¿$

$P(M \cap A) = ¿$

$E_{AB} = ¿$

$E_{11} = ¿$

$E_{11} = ¿$

	Party A	Party B	Row Total
Male			120
Female			80
Col Total	105	95	200

Observed Frequency vs Expected Frequency

=Observed frequency in ith row and jth column

=Expected frequency in ith row and jth column.

=

Observed Frequency	Party A	Party B	Row Total
Male	55	65	120 (M)
Female	50	30	80
Col Total	105 (A)	95	200 (N)

$$P(M)=M/N=120/200=0.6, P(F)=0.4$$

$$P(A)=A/N=105/200=0.525, P(B)=0.475$$

$$P(A \cap B)=P(A)\times P(B)$$

$$P(M \cap A)=P(M)\times P(A)=0.6\times 0.525=0.315$$

$$E_{AB}=P(A \cap \hat{B})\times N$$

$$E_{11}=P(M \cap A)\times N=0.315\times 200=63$$

	Party A	Party B	Row Total
Male			120
Female			80
Col Total	105	95	200

$$E_{11}=P(M \cap A)\times N=P(M)\times P(A)\times N=\frac{M}{N}\times\frac{A}{N}\times N=\frac{M\times A}{N}=\frac{120\times 105}{200}=63$$

Calculation of

Number of Independent value which are assigned to statistical distribution

(n-1) or (r-1)(c-1)

Observed Frequency	Party A	Party B	Row Total
Male	55	65	120
Female	50	30	80
Col Total	105	95	200

	Party A	Party B	Row Total
Male			120
Female			80
Col Total	105	95	200

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

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

x^2 Table

Degrees of freedom	Probability under H_0 that of χ^2 > Chi square						
	.99	.95	.50	.10	.05	.02	.01
1	.000157	.00393	.455	2.706	3.841	5.412	6.635
2	.0201	.103	1.386	4.605	5.991	7.824	9.210
3	.115	.352	2.366	6.251	7.815	9.837	11.341
4	.297	.711	3.357	7.779	9.488	11.668	13.277
5	.554	.1145	4.351	9.236	11.070	13.388	15.086
6	.872	1.635	5.348	10.645	12.592	15.033	16.812
7	1.239	2.167	6.346	12.017	14.067	16.622	18.475
8	1.646	2.733	7.344	13.362	15.507	18.168	20.090
9	2.088	3.325	8.343	14.684	16.919	19.679	21.666
10	2.558	3.940	9.342	15.987	18.307	21.161	23.209
11	3.053	4.575	10.341	17.275	19.675	22.618	24.725
12	3.571	5.226	11.340	18.549	21.026	24.054	26.217
13	4.107	5.892	12.340	19.812	22.362	25.472	72.688
14	4.660	6.571	13.339	21.064	23.685	26.873	29.141
15	4.229	7.261	14.339	22.307	24.996	28.259	30.578

Degrees of freedom	Probability under H_0 that of χ^2 > Chi square						
	.99	.95	.50	.10	.05	.02	.01
16	5.812	7.962	15.338	23.542	26.296	29.633	32.000
17	6.408	8.672	16.338	24.769	27.587	30.995	33.409
18	7.015	9.390	17.338	25.989	28.869	32.346	34.805
19	7.633	10.117	18.338	27.204	30.144	33.687	36.191
20	8.260	10.851	19.337	28.412	31.410	35.020	37.566
21	8.897	11.591	20.337	29.615	32.671	36.343	38.932
22	9.542	12.338	21.337	30.813	33.924	37.659	40.289
23	10.196	13.091	22.337	32.007	35.172	38.968	41.638
24	10.856	13.848	23.337	32.196	36.415	40.270	42.980
25	11.524	14.611	24.337	34.382	37.652	41.566	44.314
26	12.198	15.379	25.336	35.363	38.885	41.856	45.642
27	12.879	16.151	26.336	36.741	40.113	44.140	46.963
28	13.565	16.928	27.336	37.916	41.337	45.419	48.278
29	14.256	17.708	28.336	39.087	42.557	46.693	49.588
30	14.953	18.493	29.336	40.256	43.773	47.962	50.892

Problem - 1

A die is thrown 132 times with following results: Is the die biased?

λ

λ

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

Is the die unbiased?

Answer:Problem – 1

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:

Table 10.2

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

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

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.

Problem - 2

2. Find the value of X^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

Class	Obs Freq	Exp Freq	Oi – Ei	(Oi – Ei)^2/Ei
A&B				
C				
D&E				

Answer: Problem - 2

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 :

Table 10.3

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

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

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

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

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

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 - 4

Eight coins were tossed 256 times and the following results were obtained:

Numbers of heads	0	1	2	3	4	5	6	7	8
Frequency	2	6	30	52	67	56	32	10	1

Are the coins biased? Use χ^2 test.

Class (heads)	Exp Freq
0	
1	
2	
3	
4	
5	
6	
7	
8	

Class (heads)	Obs Freq	Exp Freq	Oi – Ei	(Oi – Ei)^2/Ei
0	2			
1	6			
2	30			
3	52			
4	67			
5	56			
6	32			
7	10			
8	1			

Answer: Problem - 4

Solution: Let us take the hypothesis that the coins are not biased. If that is so, the probability of any one coin falling with head upward is 1/2 and with tail upward is 1/2 and it remains the same whatever be the number of throws. In such a case the expected values of getting 0, 1, 2, … heads in a single throw in 256 throws of eight coins will be worked out as follows*.

Table 10.7

Events or No. of heads	*Expected frequencies*
0	${}^8C_0\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^8 \times 256 = 1$
1	${}^8C_1\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^7 \times 256 = 8$
2	${}^8C_2\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^6 \times 256 = 28$
3	${}^8C_3\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^5 \times 256 = 56$
4	${}^8C_4\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right)^4 \times 256 = 70$
5	${}^8C_5\left(\frac{1}{2}\right)^5\left(\frac{1}{2}\right)^3 \times 256 = 56$
6	${}^8C_6\left(\frac{1}{2}\right)^6\left(\frac{1}{2}\right)^2 \times 256 = 28$
7	${}^8C_7\left(\frac{1}{2}\right)^7\left(\frac{1}{2}\right)^1 \times 256 = 8$
8	${}^8C_8\left(\frac{1}{2}\right)^8\left(\frac{1}{2}\right)^0 \times 256 = 1$

The value of χ^2 can be worked out as follows:

Answer: Problem - 4

Table 10.8

No. of heads	Observed frequency O_i	Expected frequency E_i	$O_i - E_i$	$(O_i - E_i)^2/E_i$
0	2	1	1	1/1 = 1.00
1	6	8	–2	4/8 = 0.50
2	30	28	2	4/28 = 0.14
3	52	56	–4	16/56 = 0.29
4	67	70	–3	9/70 = 0.13
5	56	56	0	0/56 = 0.00
6	32	28	4	16/28 = 0.57
7	10	8	2	4/8 = 0.50
8	1	1	0	0/1 = 0.00

$$\therefore \quad \chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 3.13$$

$\therefore$ Degrees of freedom = $(n - 1) = (9 - 1) = 8$

The table value of χ^2 for eight degrees of freedom at 5 per cent level of significance is 15.507.

The calculated value of χ^2 is much less than this table and hence it is insignificant and can be scribed due to fluctuations of sampling. The result, thus, supports the hypothesis and we may say hat the coins are not biased.

Problem - 5

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

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 14.66$$

$$df = (r-1)(c-1) = (2-1)(2-1) = 1$$

Critcal value of for df=1,
at 5% level of significance is 3.841
Computed (=14.66) > 3.841.
So reject H0 and conclude that Vaccine is effective.

Problem 6 – Star Trek: fatality vs shirt color

	Blue	Gold	Red	Row total
Dead	7	9	24	40
Alive	129	46	215	390
Column total	136	55	239	N = 430
Column percentage (Dead)	5.15%	16.36%	10.4%	

H0 : fatality and shirt color are related

Ha : fatality is not related to shirt color.

Exp Freq	Blue(B)	Gold(G)	Red®	Row Tol
Dead(D)			22.23	40
Alive(A)	123.35	49.88	216.77	390
Col Total	136	55	239	430

Uniform	Status	Observed	Expected	Squared difference/Expected
Blue	Dead	7	12.65	2.52
Blue	Alive	129	123.35	0.26
Gold	Dead	9	5.12	2.94
Gold	Alive	46	49.88	0.30
Red	Dead	24	22.3	0.13
Red	Alive	215	216.77	0.01
			Sum	6.17

@ 5% significance level

Since = 6.17 > 5.991

Evidence to Reject H0 and Accept Ha

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 2

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).

Refer text book and solve worked example problems :- 11.2 , 11.3,11.7 to 11.14.

Also solve exercise problems :- 3, 4, 5, 6, 7.

Introduction - revisit

- Principal Instrument of research
  - Function is to suggest experiments and observation.
  - Hypothesis testing is often used strategy for deciding whether sample data offer such support for hypothesis that generalization can be made.

THANK YOU

Raghu B.A
Priya B.
Santhosh Kumar V.
Department of Computer Science & Engineering

RESEARCH METHODOLOGY

UE20CS506A

Unit-03:

Testing of hypotheses and Data Analysis

Raghu B. A
Priya B.
Santhosh Kumar V
Department of Computer Science & Engineering

RESEARCH METHODOLOGY

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

Santhosh Kumar V
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:

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

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

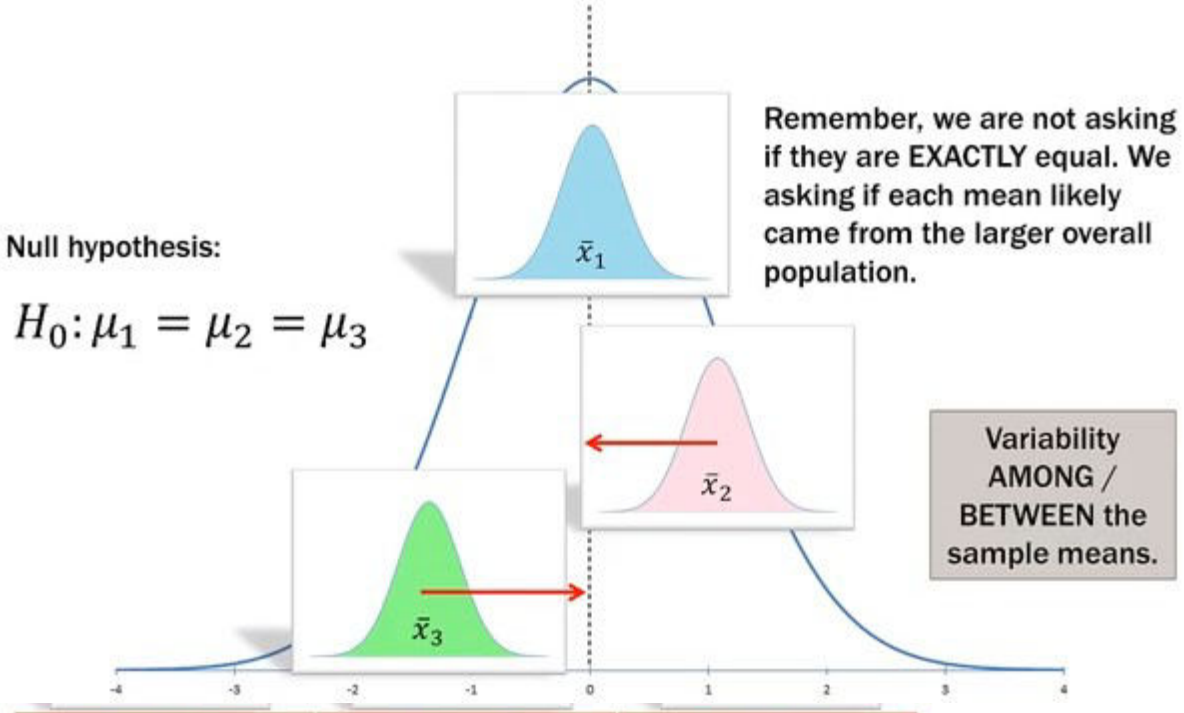
In such a situation t-test is used and the test statistic t is worked out as under:

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

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

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

ANOVA

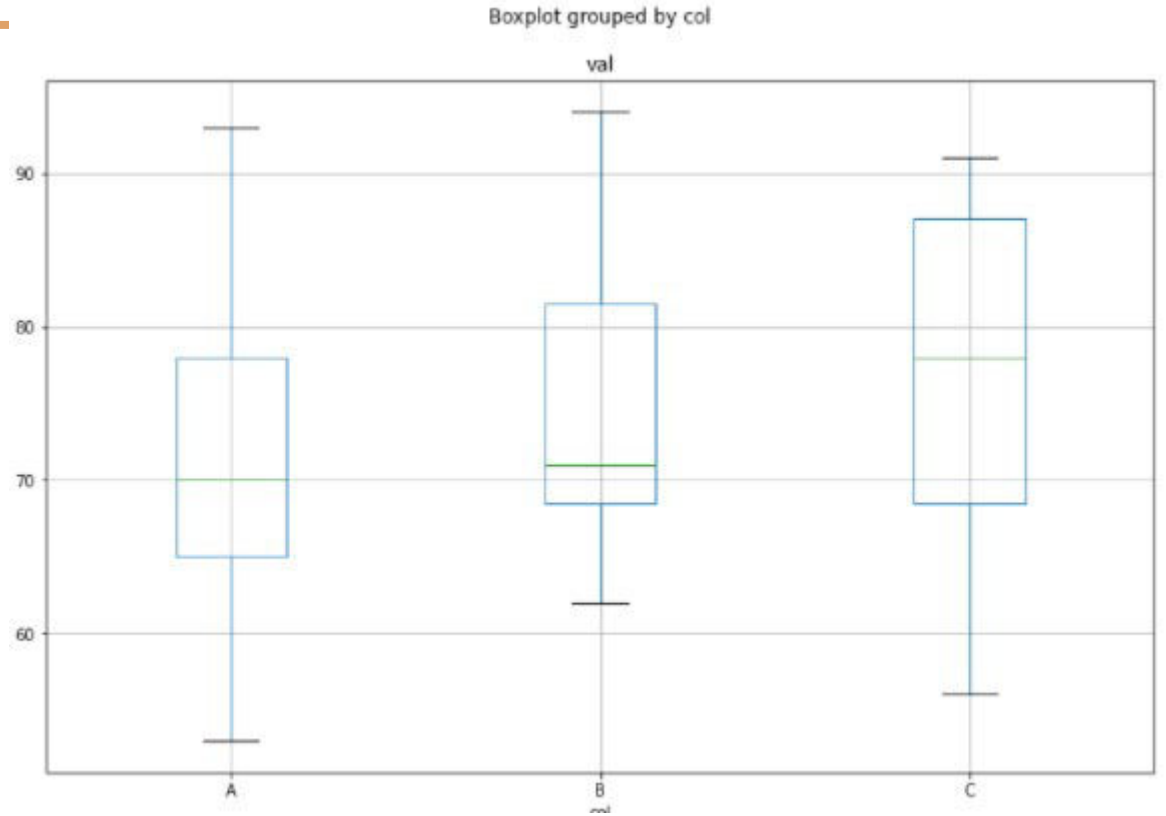


Year 1 Scores	Year 2 Scores	Year 3 Scores
82	71	64
93	62	73
61	85	87
74	94	91
69	78	56
70	66	78
53	71	87
$\bar{x}_1 = 71.71$	$\bar{x}_2 = 75.29$	$\bar{x}_3 = 76.57$

Overall Mean:

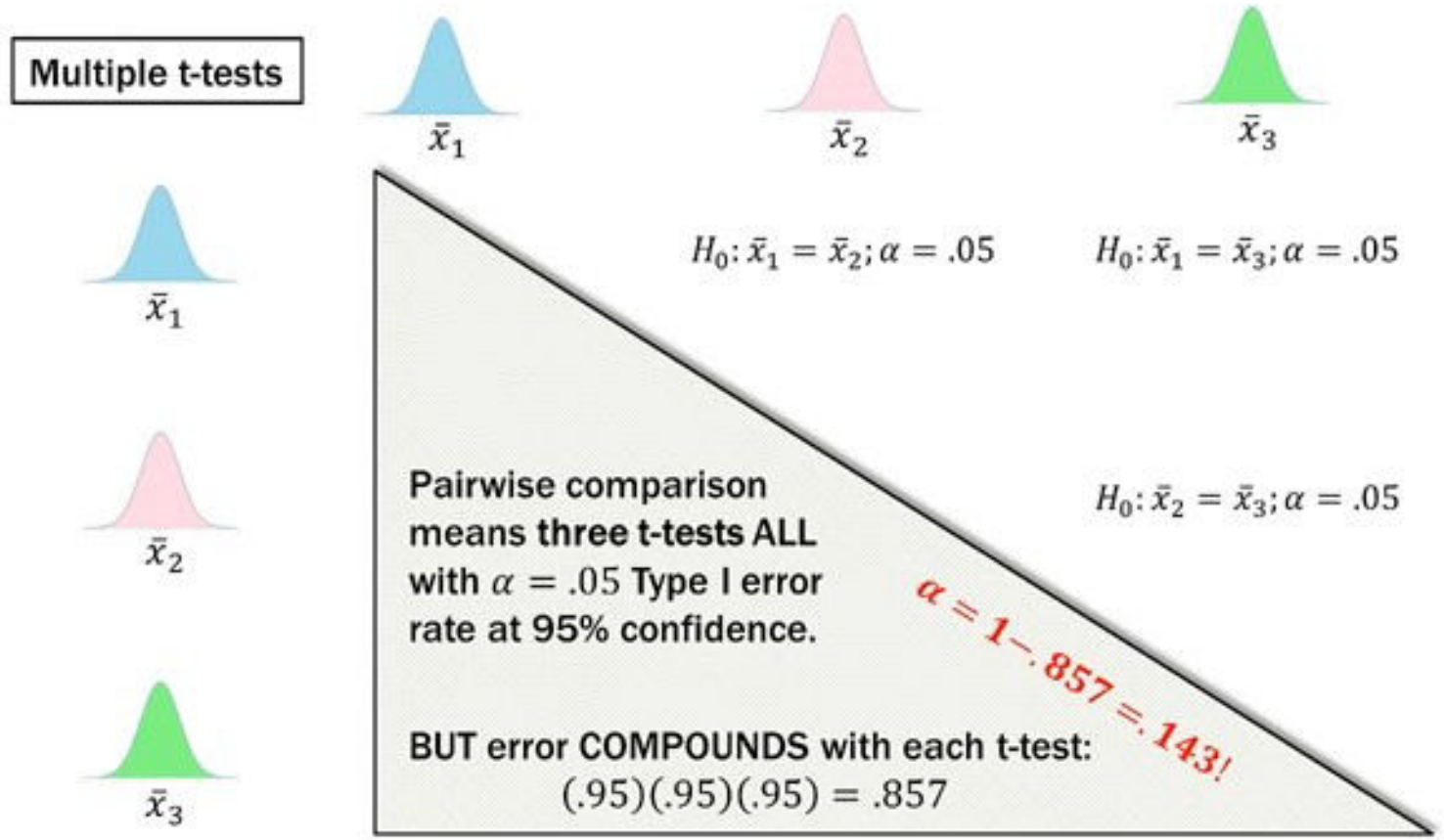
The mean of all 21 scores taken together.

$$\bar{\bar{x}} = 74.52$$

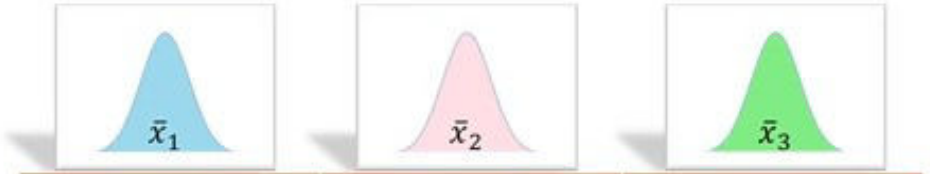


Ref: statistics 101: Anova visual

Multiple t-test (not a solution)



ANOVA

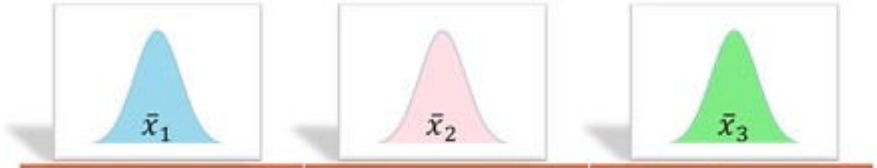


Year 1 Scores	Year 2 Scores	Year 3 Scores
82	71	64
93	62	73
61	85	87
74	94	91
69	78	56
70	66	78
53	71	87
$\bar{x}_1 = 71.71$	$\bar{x}_2 = 75.29$	$\bar{x}_3 = 76.57$

Overall Mean:

The mean of all 21 scores taken together.

$$\bar{\bar{x}} = 74.52$$



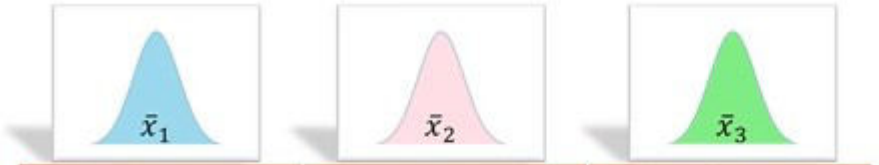
Year 1 Scores	Year 2 Scores	Year 3 Scores
82	71	64
93	62	73
61	85	87
74	94	91
69	78	56
70	66	78
53	71	87
$\bar{x}_1 = 71.71$	$\bar{x}_2 = 75.29$	$\bar{x}_3 = 76.57$

SST
(total / overall) sum of squares

1. Find difference between each data point and the overall mean.
2. Square the difference.
3. Add them up

$$\bar{\bar{x}} = 74.52$$

ANOVA



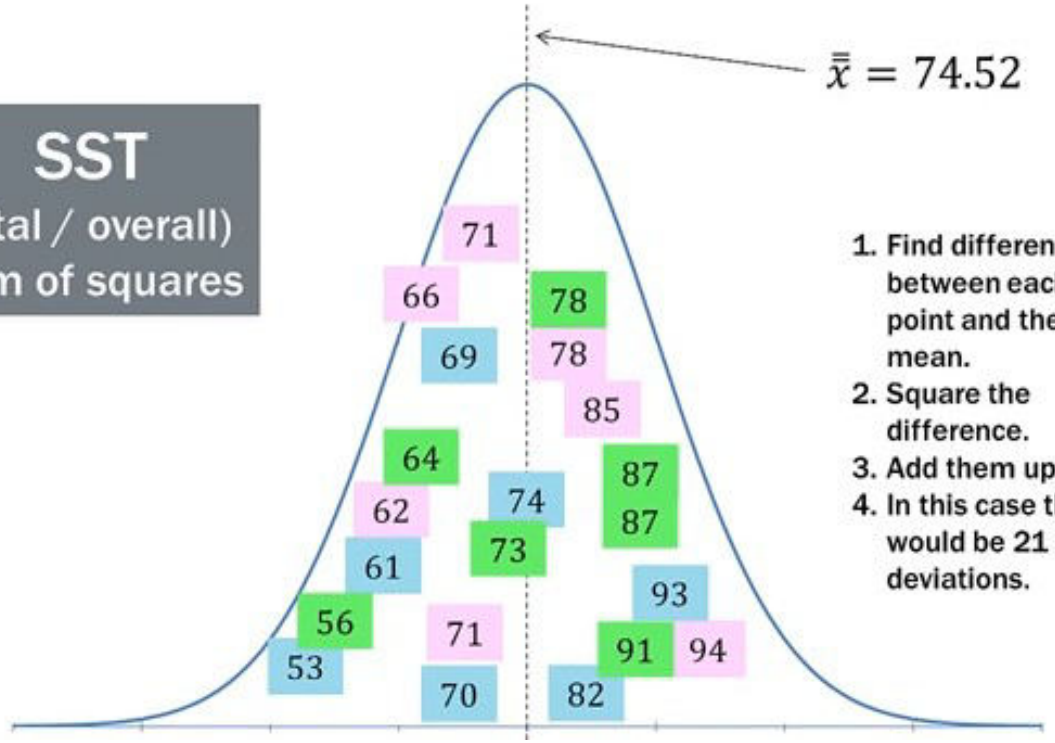
Year 1 Scores	Year 2 Scores	Year 3 Scores
82	71	64
93	62	73
61	85	87
74	94	91
69	78	56
70	66	78
53	71	87
$x_1 = 71.71$	$x_2 = 75.29$	$x_3 = 76.57$

SST
(total / overall)
sum of squares

1. Find difference between each data point and the overall mean.
2. Square the difference.
3. Add them up

$\bar{\bar{x}} = 74.52$

SST
(total / overall)
sum of squares

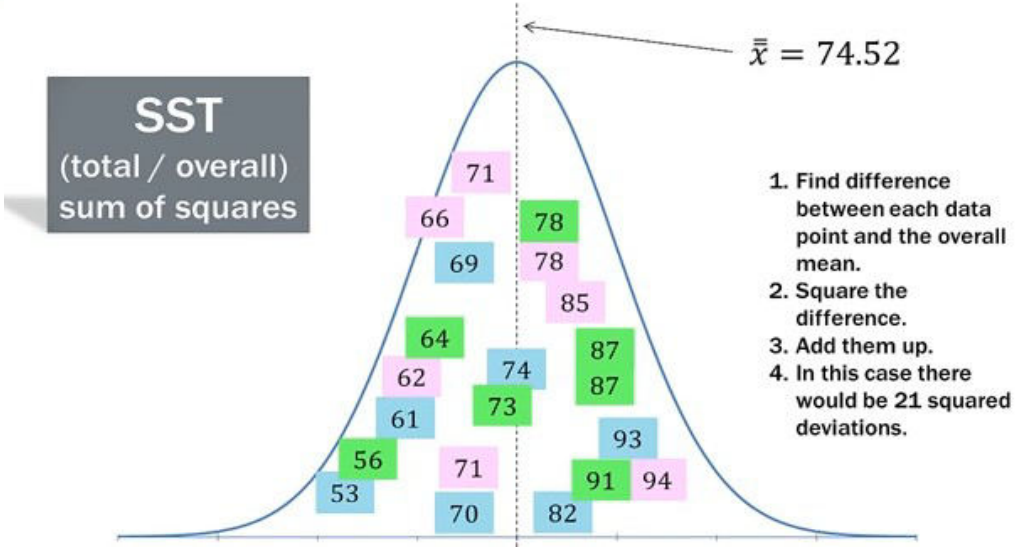


1. Find difference between each data point and the overall mean.
2. Square the difference.
3. Add them up.
4. In this case there would be 21 squared deviations.

ANOVA

SST
(total / overall)
sum of squares

1. Find difference between each data point and the overall mean.
2. Square the difference.
3. Add them up

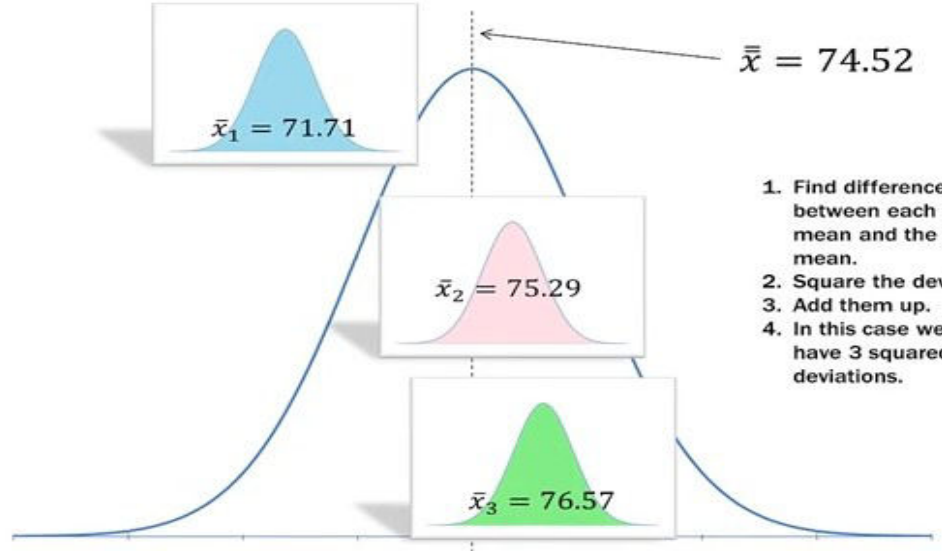


SST
(total / overall)
sum of squares

1. Find difference between each data point and the overall mean.
2. Square the difference.
3. Add them up.
4. In this case there would be 21 squared deviations.

SSC
(column/ between)
sum of squares

1. Find difference between each group mean and the overall mean.
2. Square the deviations.
3. Add them up.
4. In this case we would have 3 squared deviations.



1. Find difference between each sample mean and the overall mean.
2. Square the deviations.
3. Add them up.
4. In this case we would have 3 squared deviations.

ANOVA

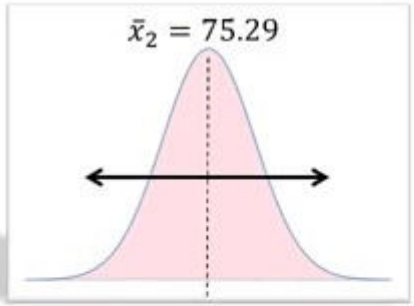
SSE
(within / error)
sum of squares

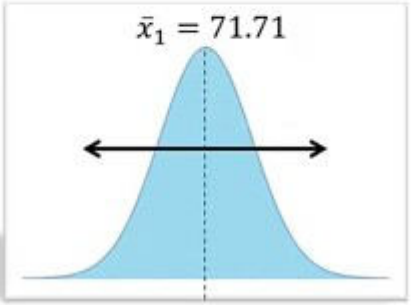
$\bar{\bar{x}} = 74.52$

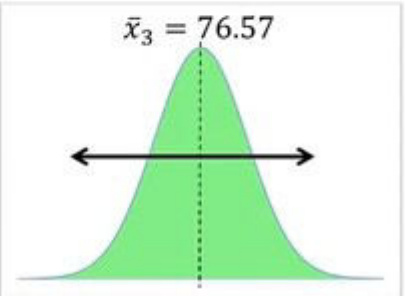
1. Find difference between each data point and its **column** mean.
2. Square each deviation.
3. Add them up the squared deviations.
4. In this case we would have 21 squared deviations.

SSE
(within / error)
sum of squares

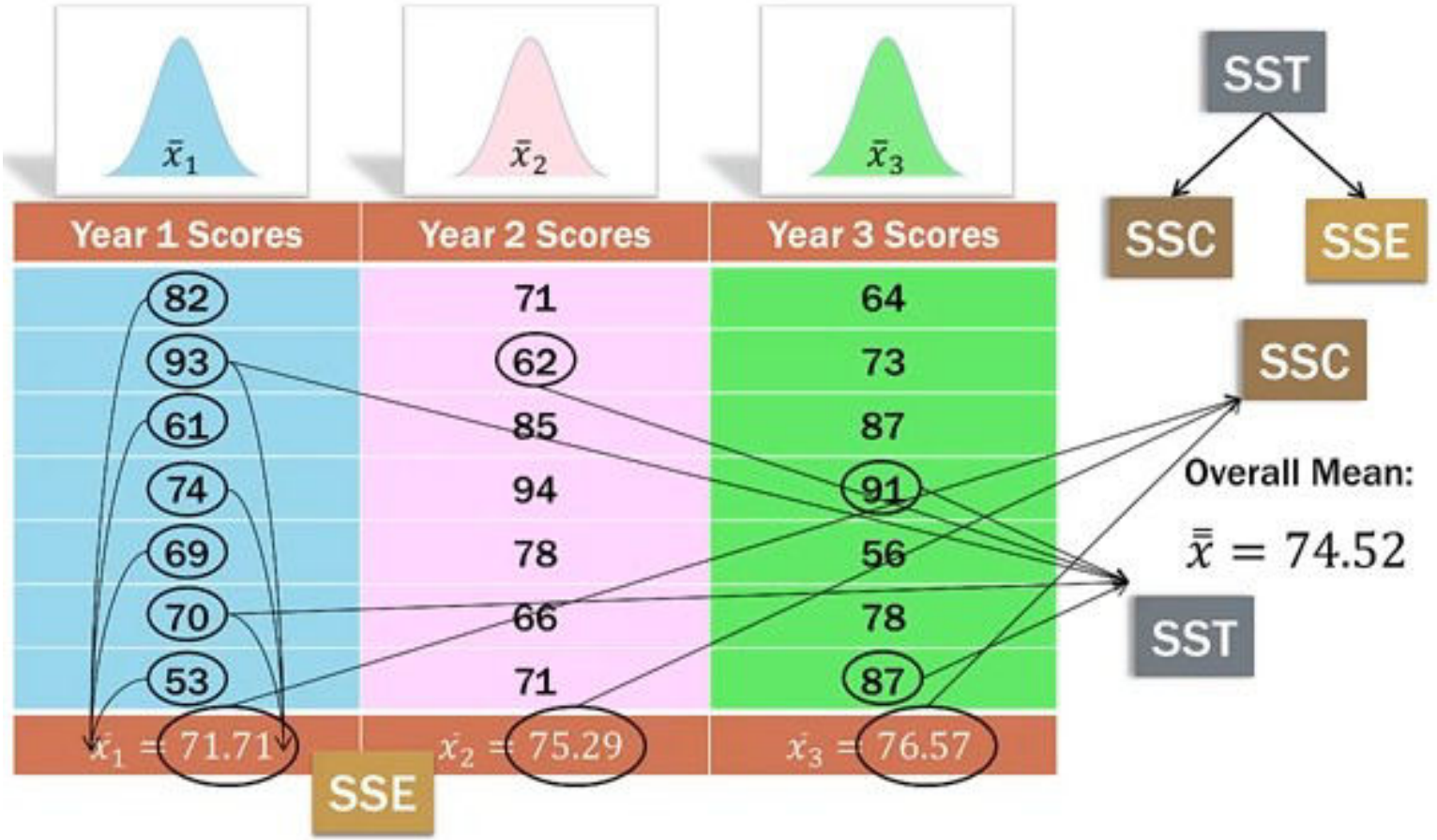
1. Find difference between each data point and its **column** mean.
2. Square each deviation.
3. Add them up the squared deviations.
4. In this case we would have 21 squared deviations.







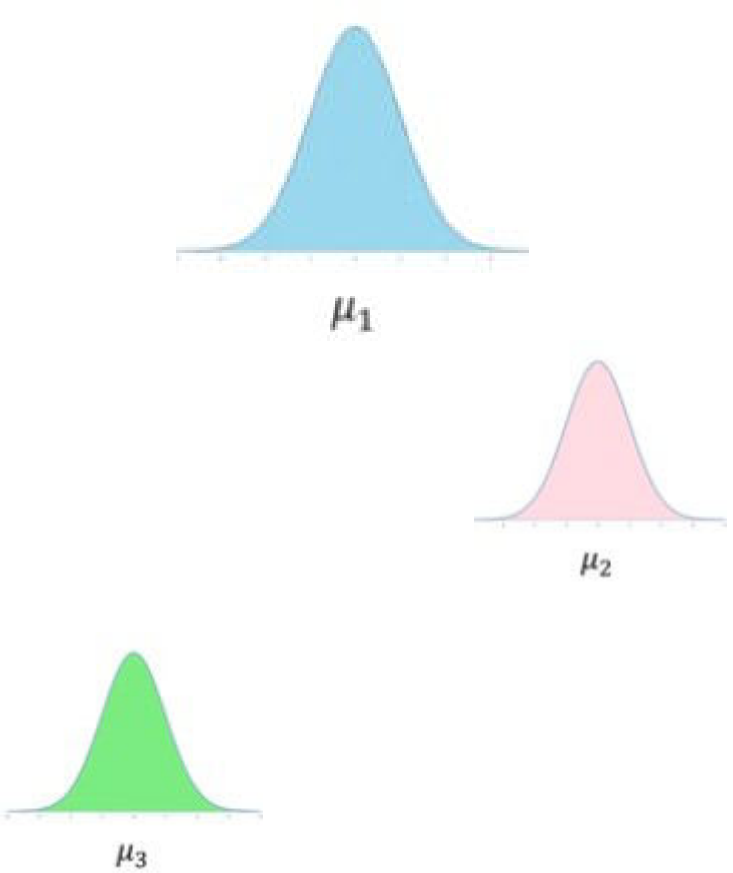
ANOVA


$\bar{x}_1$
$\bar{x}_2$
$\bar{x}_3$
Year 1 Scores
Year 2 Scores
Year 3 Scores
82
93
61
74
69
70
53
71
62
85
94
78
66
71
64
73
87
91
56
78
87
$\bar{x}_1 = 71.71$
$\bar{x}_2 = 75.29$
$\bar{x}_3 = 76.57$
SSE
SST
SSC
SSE
SSC
Overall Mean:
$\bar{\bar{x}} = 74.52$
SST

ANOVA

Compare 3 population means to see if they are different

Do all the 3 means come from the same population



Is one mean so far away , it is from a different 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
	6	5	4

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

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

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

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

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

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

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

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

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
	6	5	4

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

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

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

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

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

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

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

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

$$= 32$$

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

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

$$= 4 + 0 + 4$$

$$= 8$$

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

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

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

$$= 14 + 8 + 2$$

$$= 24$$

n = total number of items in all the sample
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
	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
	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

Raghu B.A
Priya B.
Santhosh Kumar V.
Department of Computer Science & Engineering

RESEARCH METHODOLOGY

Unit-03:

Data Representation

Raghu B. A

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.

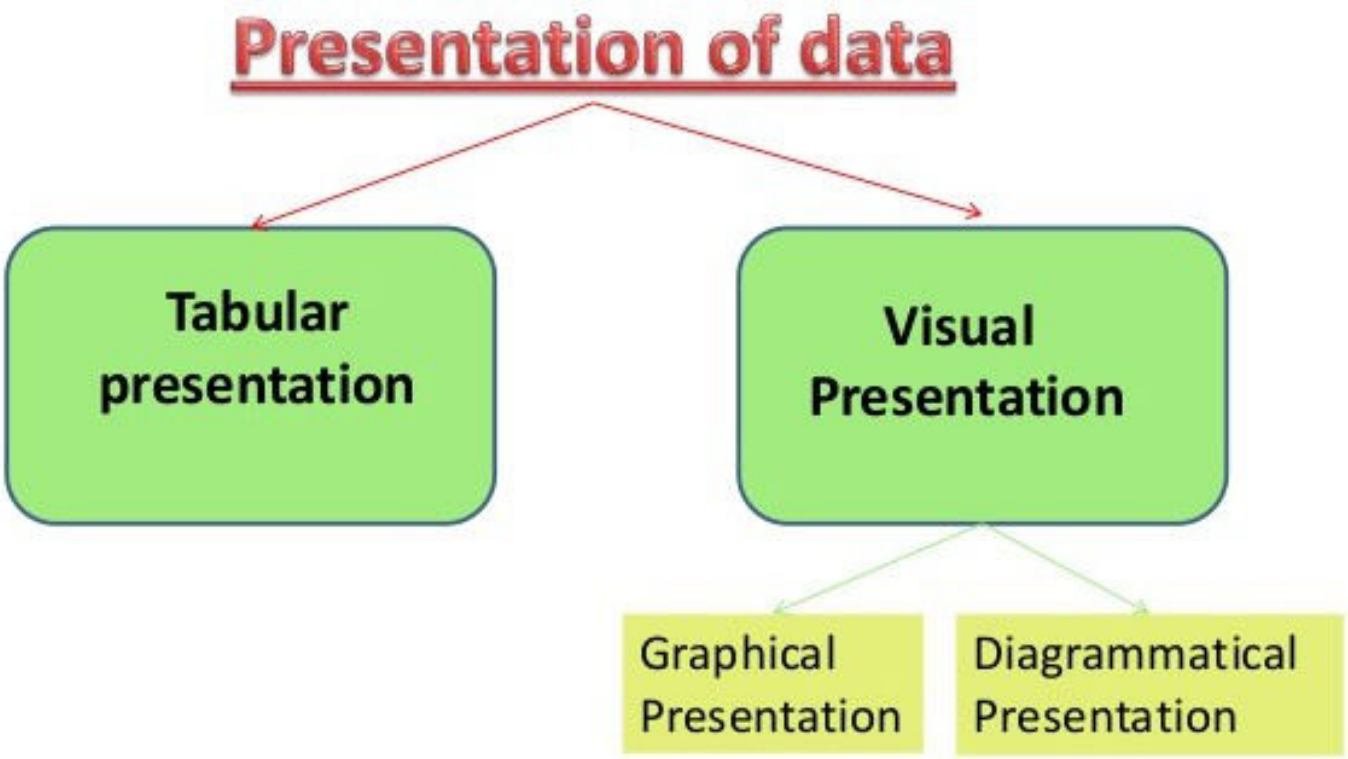
Data may be presented in(3 Methods):

- 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



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.

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

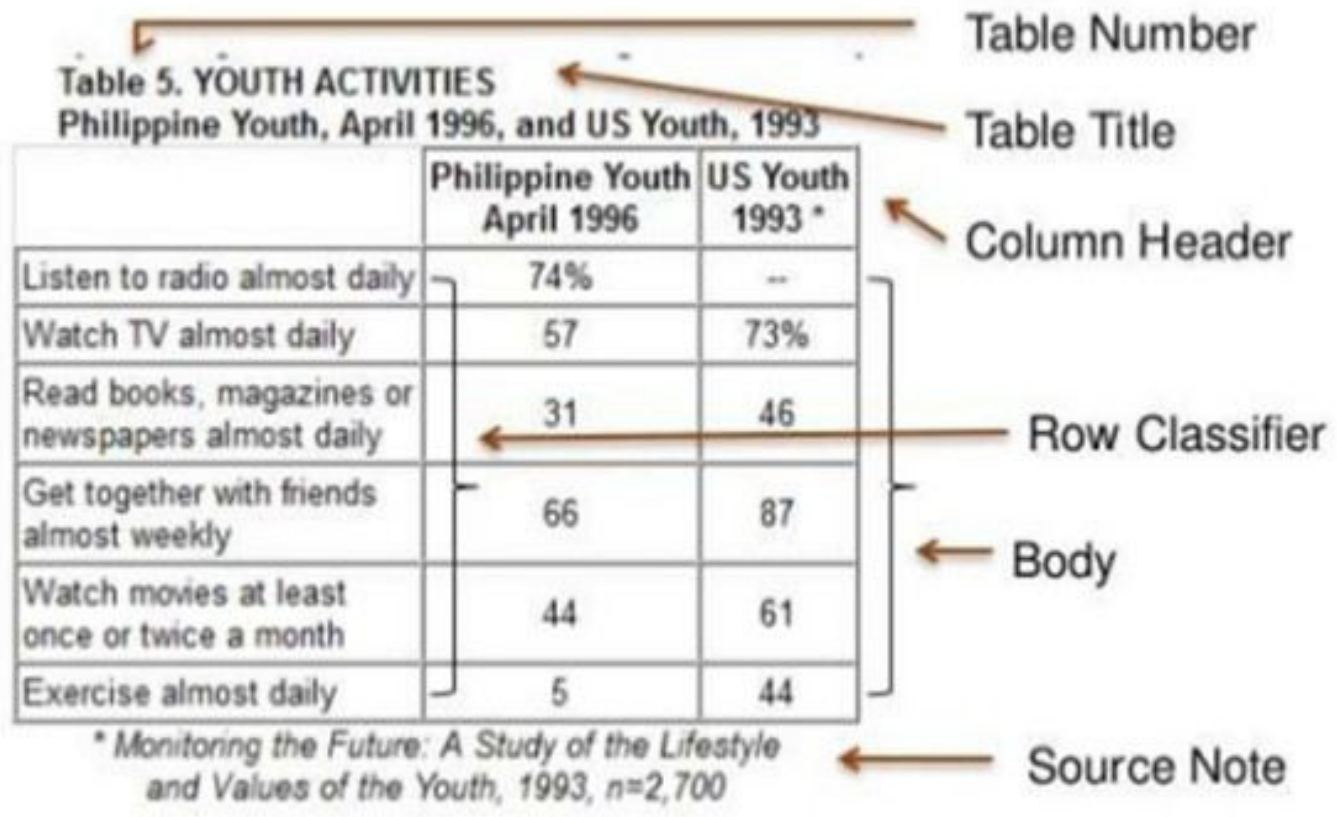
Table Number: .....
Title: .....
(Head Note, if any)

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:

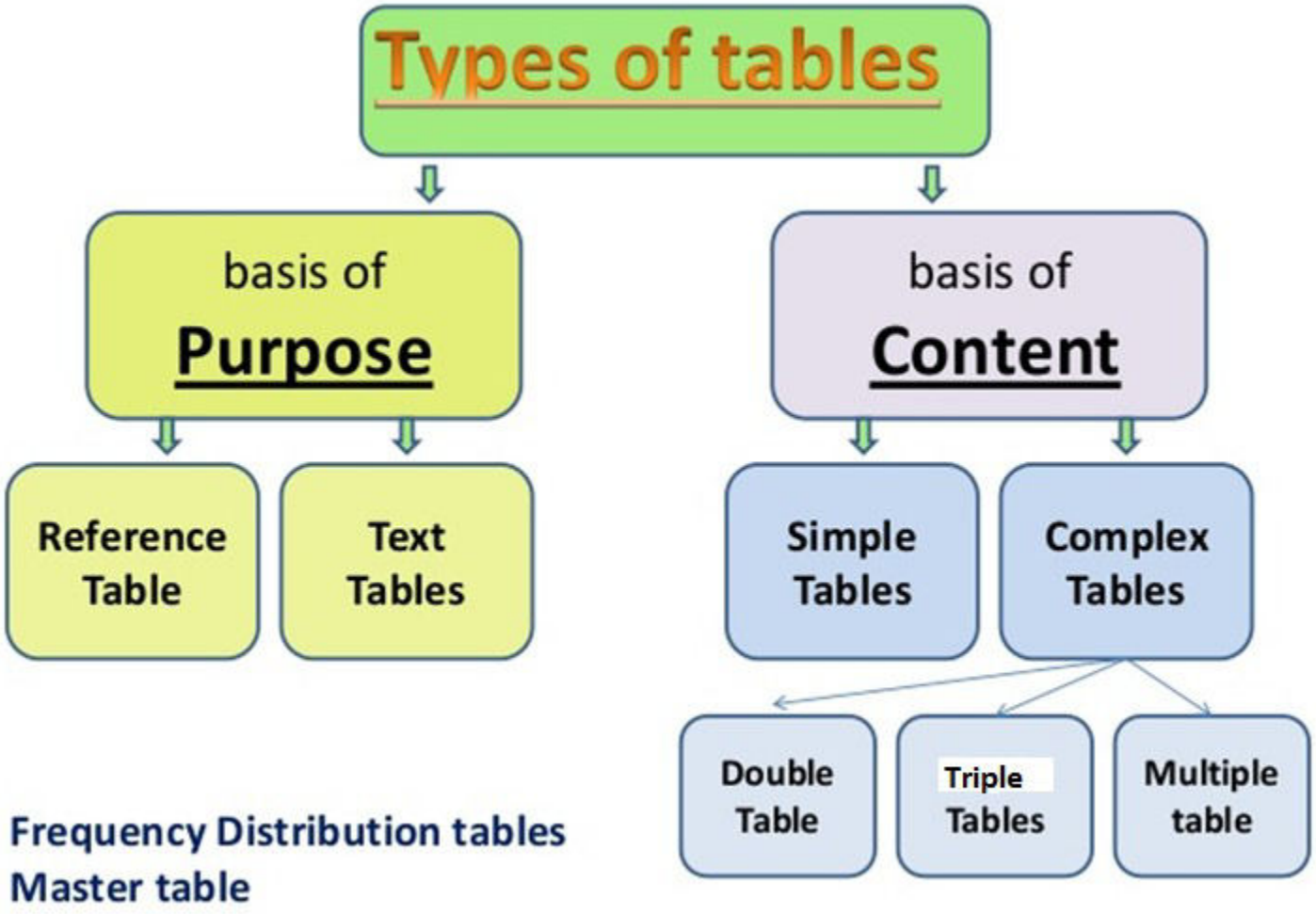


	Philippine Youth April 1996	US Youth 1993 *
Listen to radio almost daily	74%	--
Watch TV almost daily	57	73%
Read books, magazines or newspapers almost daily	31	46
Get together with friends almost weekly	66	87
Watch movies at least once or twice a month	44	61
Exercise almost daily	5	44

** 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:

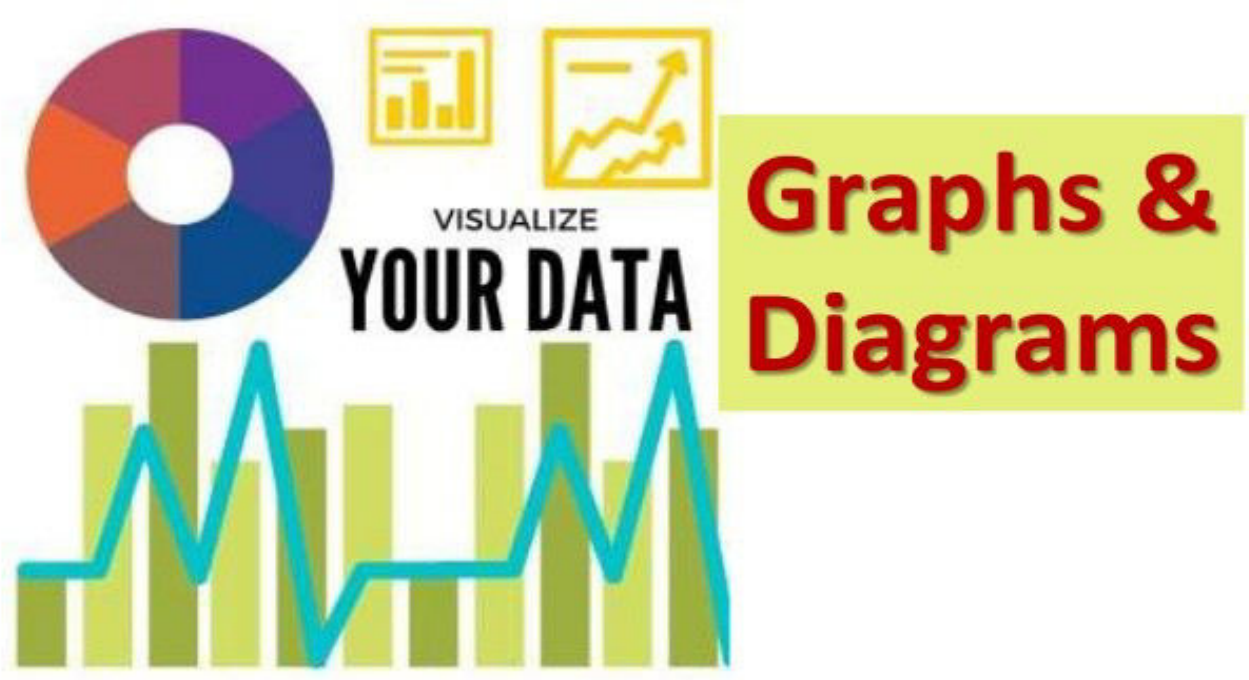
Gender		Food habit			
		Vegetarians		Non Vegetarians	
		Age <20 years	Age >=20 years	Age < 20 years	Age>= 20 years
Boys	Day scholars	0	0	1	4
	Hostellers	0	2	0	2
Girls	Day scholars	0	1	2	2
	Hostellers	1	3	8	12

Residing Area

← Age

Research Methodology

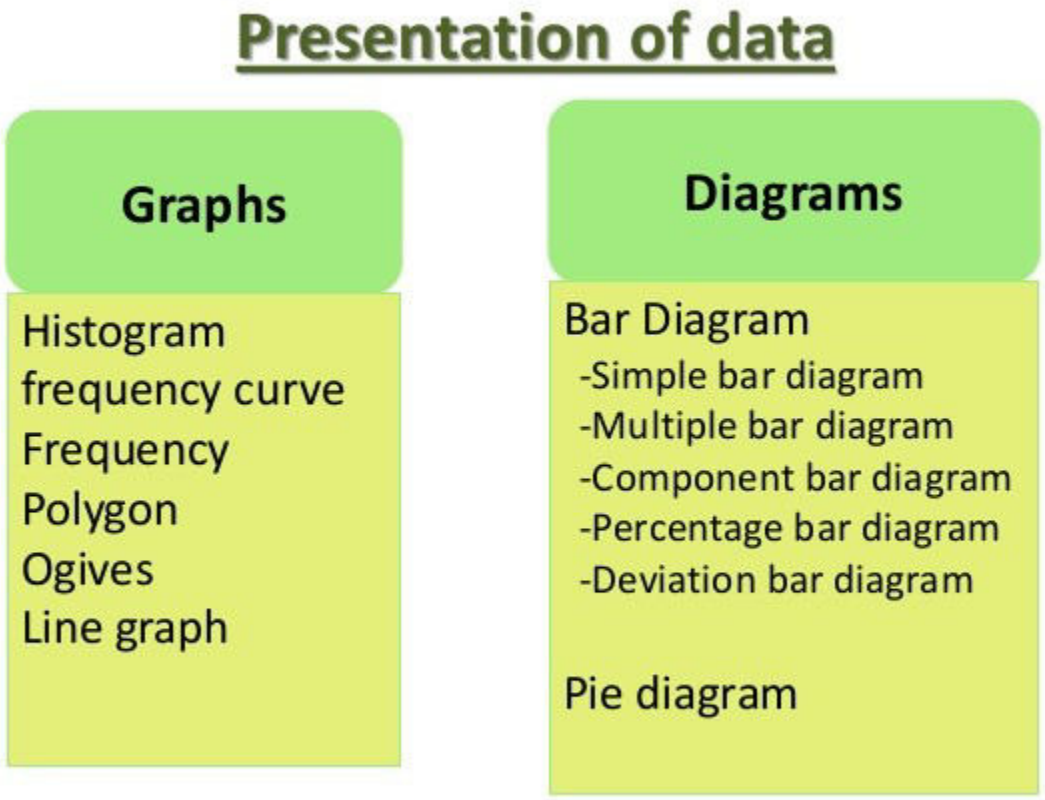
Scientific Publishing- Data representation



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:

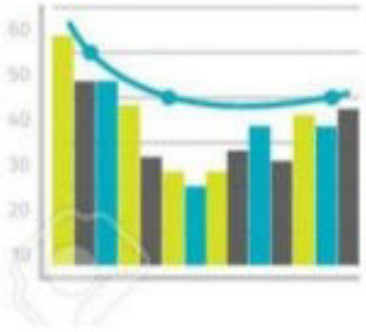
- First choose the form of diagrams /graphs which is capable of representing the given set of data.
- Title- gives information of diagrams or graphs contain.
- Scale – selection of scale should be neither too small or too large. The scale should also specify the size of unit and what it represents. (eg: 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.

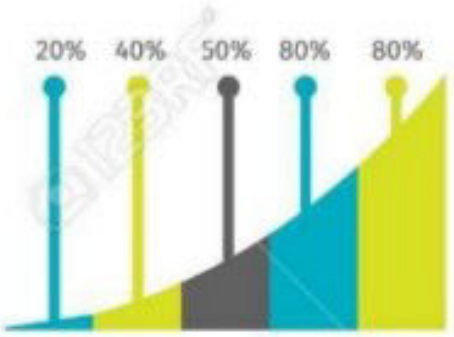
Difference between Graphs and Diagrams:

- 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

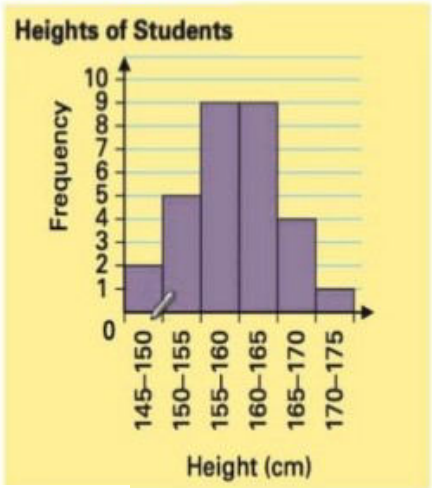


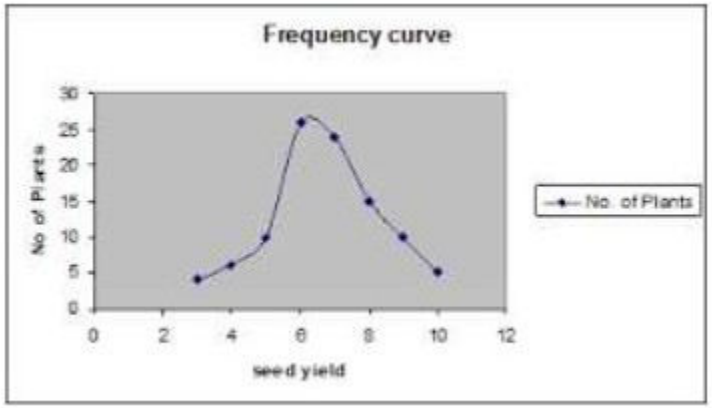


Scientific Publishing- Data representation

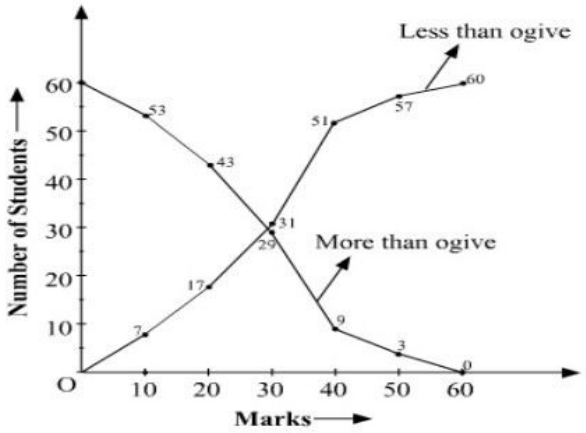
Histogram

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





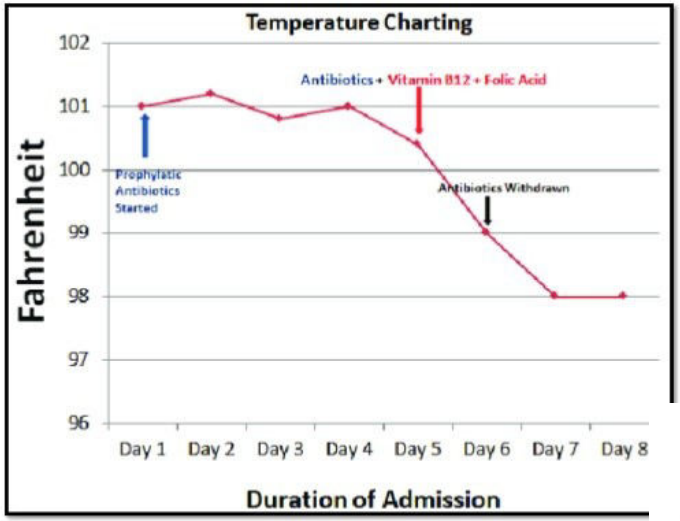
Ogives: (Cumulative Frequency Curves):



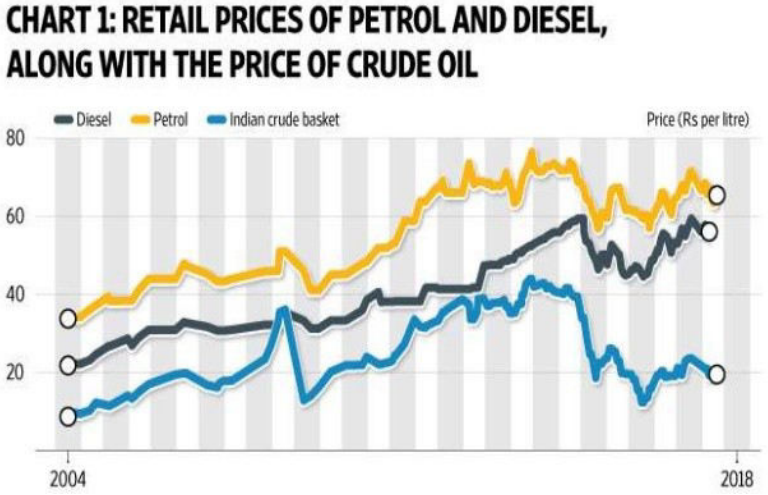
Research Methodology

Scientific Publishing- Data representation

Line Graph: (Time series graph)

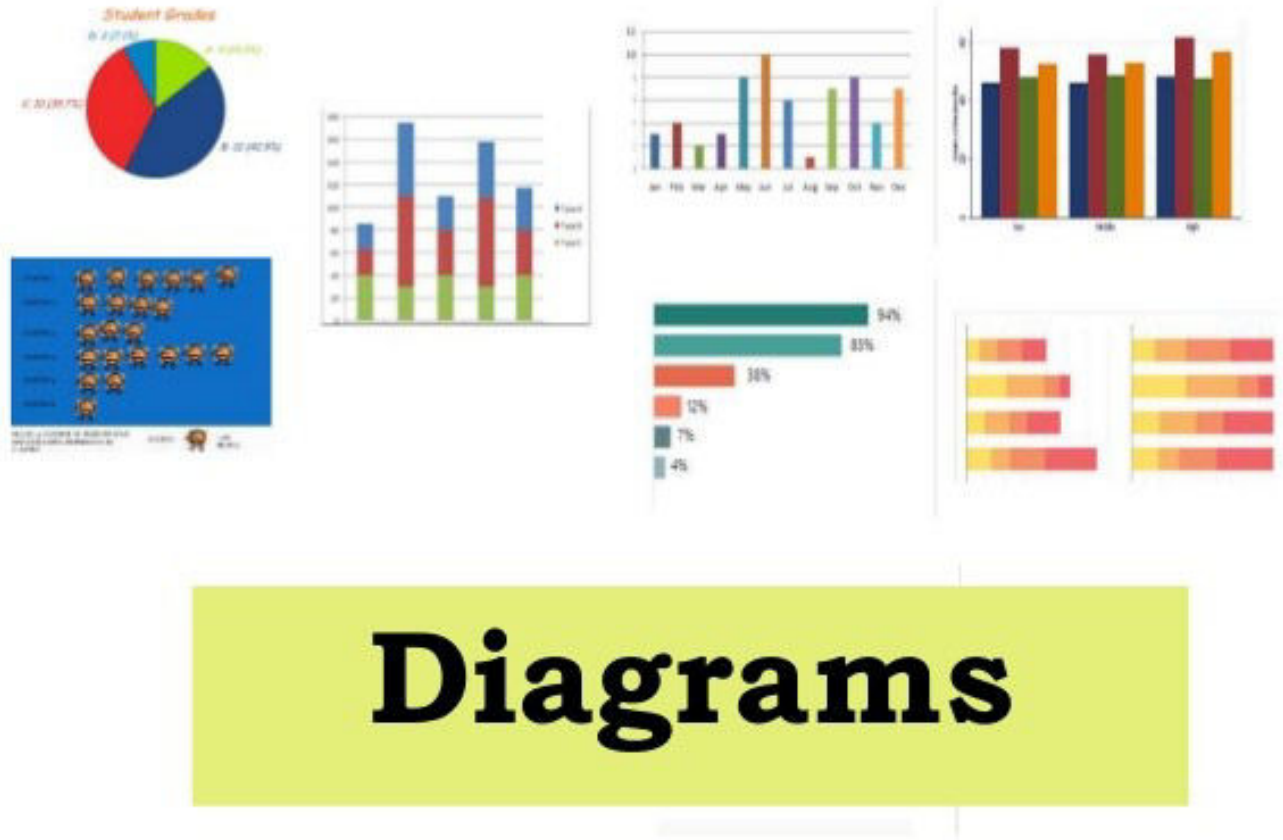


Line Graph: (Time series graph)



Research Methodology

Scientific Publishing- Data representation

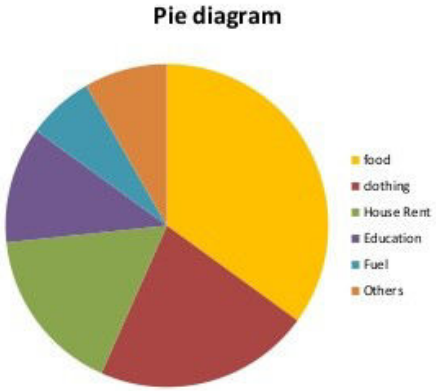


Research Methodology

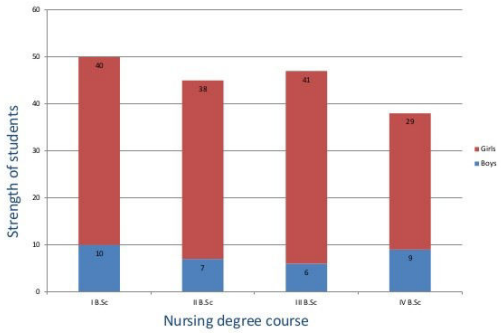
Scientific Publishing- Data representation

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

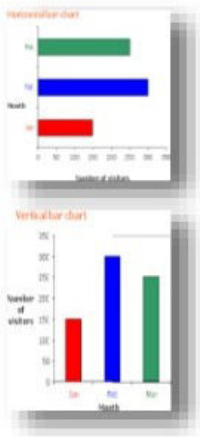
Items	Expenditure (Rs)	Degree measurement
Food	1050	126^0
Clothing	650	78^0
House rent	500	60^0
Education	350	42^0
Fuel	200	24^0
Others	250	30^0

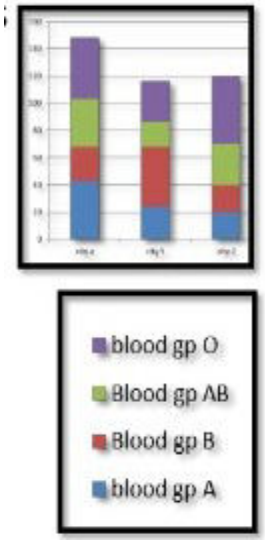


BAR DIAGRAM/ BARCHART



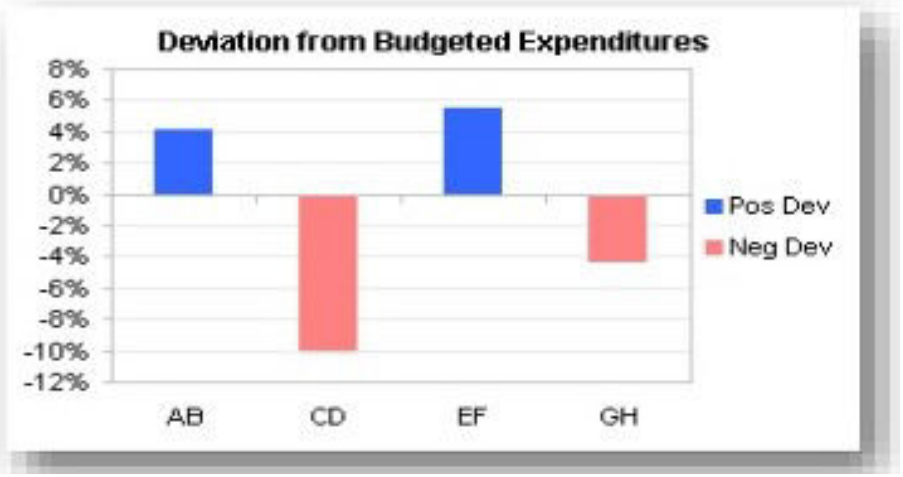
An example shows the strength of students in nursing degree course





Research Methodology

Scientific Publishing- Data representation



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

Dr. Seema Tharannum
Department of Biotechnology

seema@pes.edu

+91 80 2672 6672 Extn 352

RESEARCH METHODOLOGY

Unit-03:

Summary and Conclusions

Raghu B. A
Department of Computer Science & Engineering

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.

RESEARCH METHODOLOGY

Unit-03:

Results and Discussions

Raghu B. A

Department of Computer Science & Engineering

RESEARCH METHODOLOGY

Topic: Results and Discussion (Data Interpretation)

Department of Computer Science & Engineering

DISCUSSION

Step by step:
An effective DISCUSSION Section

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: Interpretation of findings

- This section addresses the **meaning** of your findings.
- In some cases, when your results are in the direction you predicted, this meaning was anticipated when the study was designed.
- In cases where the results are not all in the desired direction, researchers must explain why this was not the case.

  [Address sampling, measurement, and procedural issues as well as confounding variables]

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.