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.

λ

- 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".*

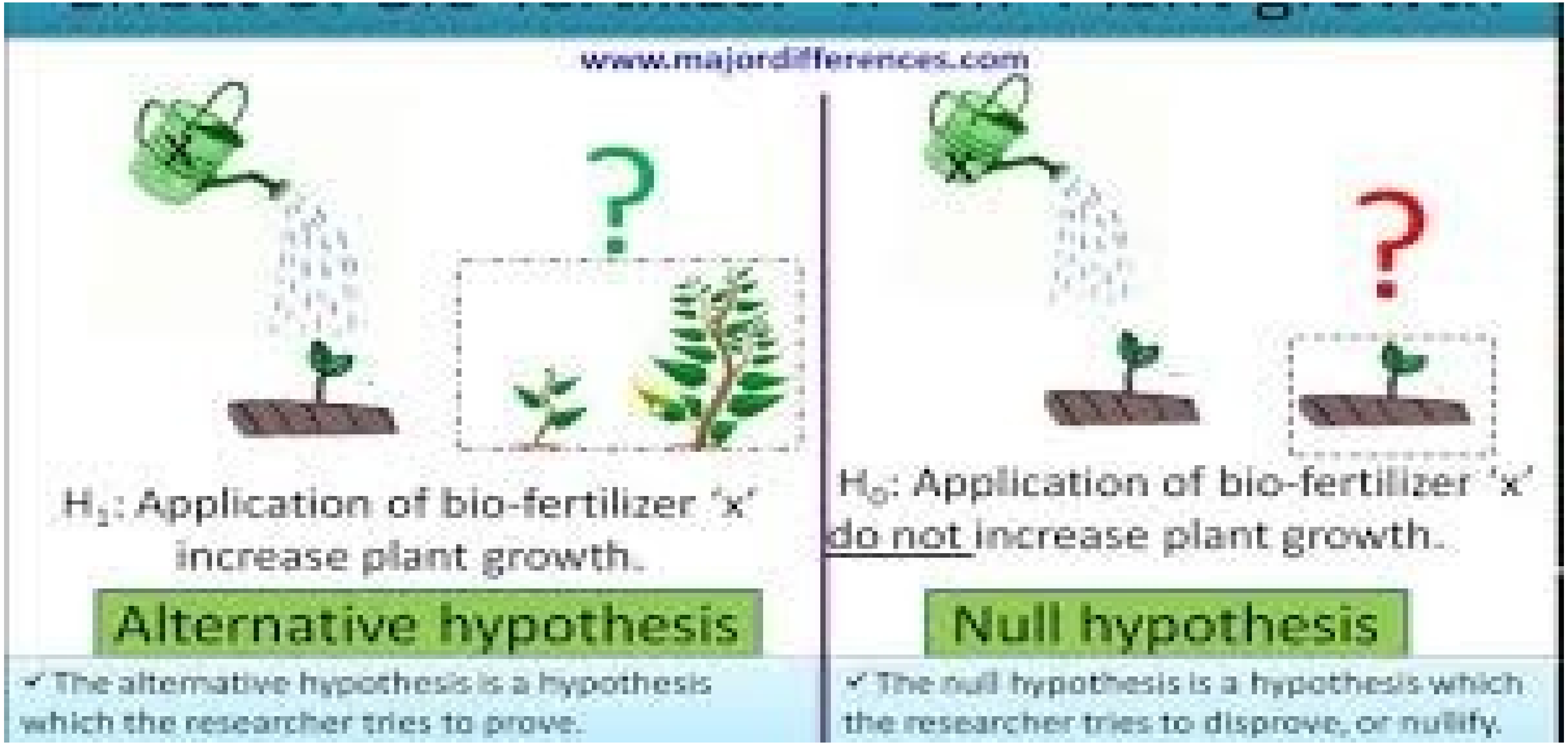
Alternate Hypothesis – If method A is superior than B (H_1).

- ***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 I Error

Type II Error

Demystifying statistics! – Lecture5

SBCM, Joint Program – Riyadh

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

Demystifying statistics! – Lecture 5

SBCM, Joint Program – Riyadh

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

Demystifying statistics! – Lecture5

SBCM, Joint Program – Riyadh

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

Demystifying statistics! – Lecture5

SBCM, Joint Program – Riyadh

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 $1 - \alpha$	Type II Error Probability β
Reject H_0	Type I Error Probability α	No Error Probability $1 - \beta$

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 $1 - \alpha$	Type II Error Probability β
Reject H_0	Type I Error Probability α	No Error Probability $1 - \beta$

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

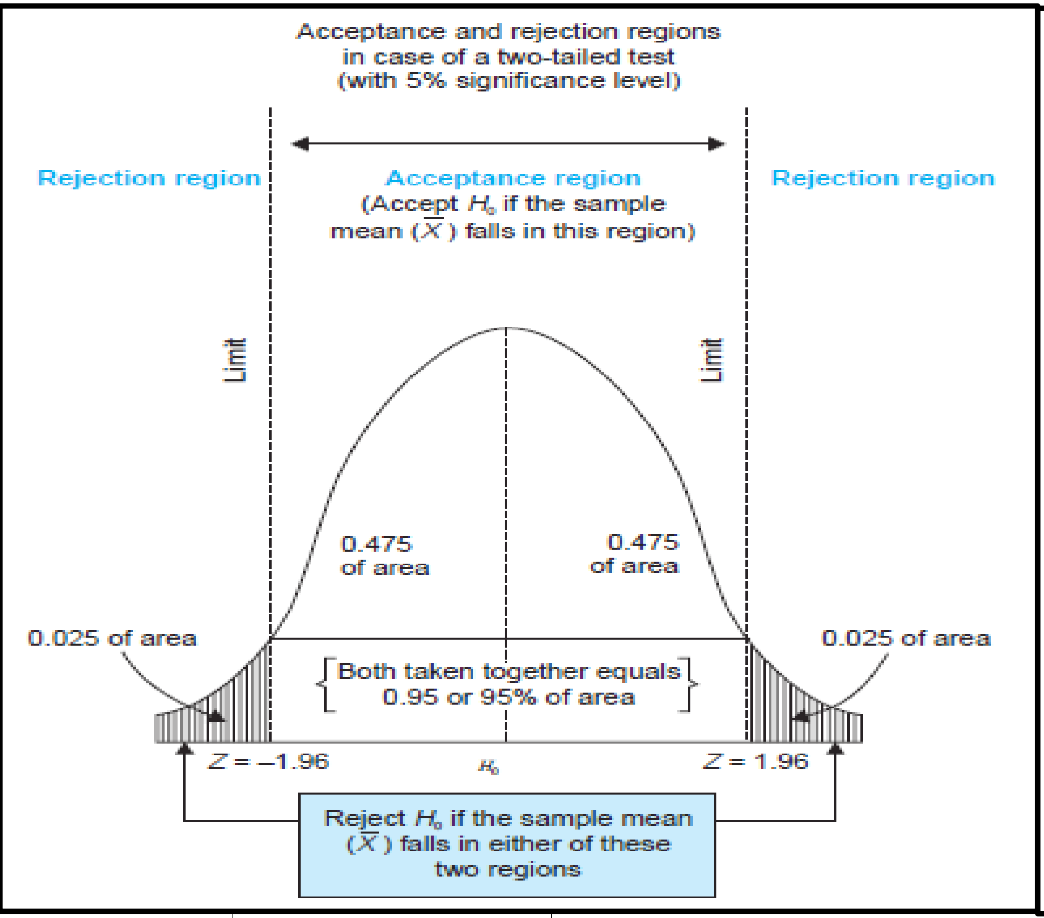
$H_0:\mu >= \mu_{H0}$ Aganist $H_a:\mu < \mu_{H0}$

If we have ≠ in alternate hypotheses – Two tailed test

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

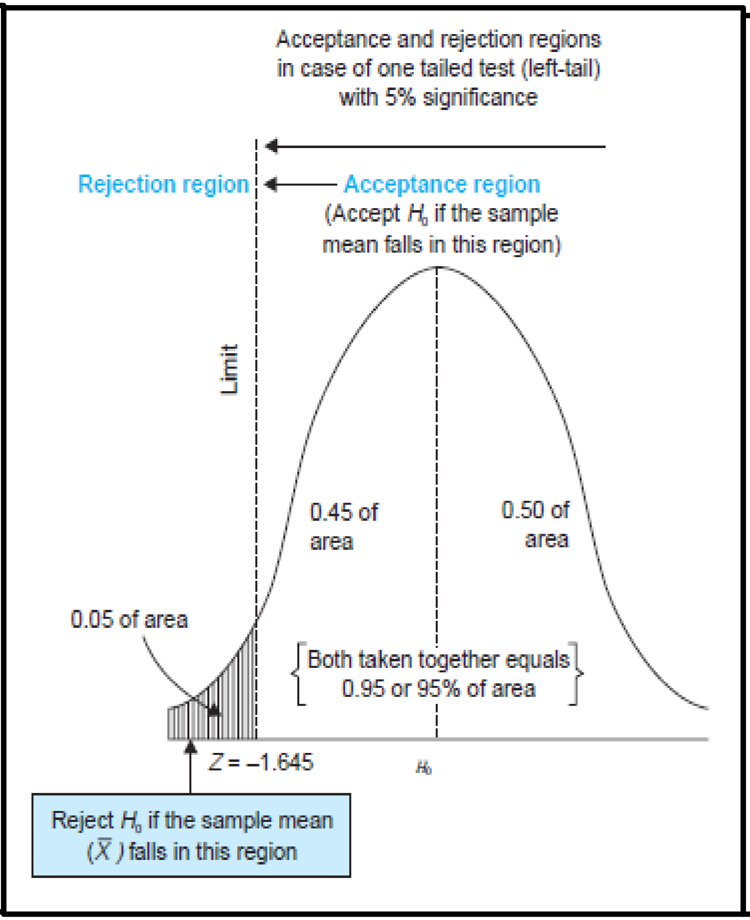
If we have < sign in alternate hypotheses – left **tailed**

One tailed and two tailed test



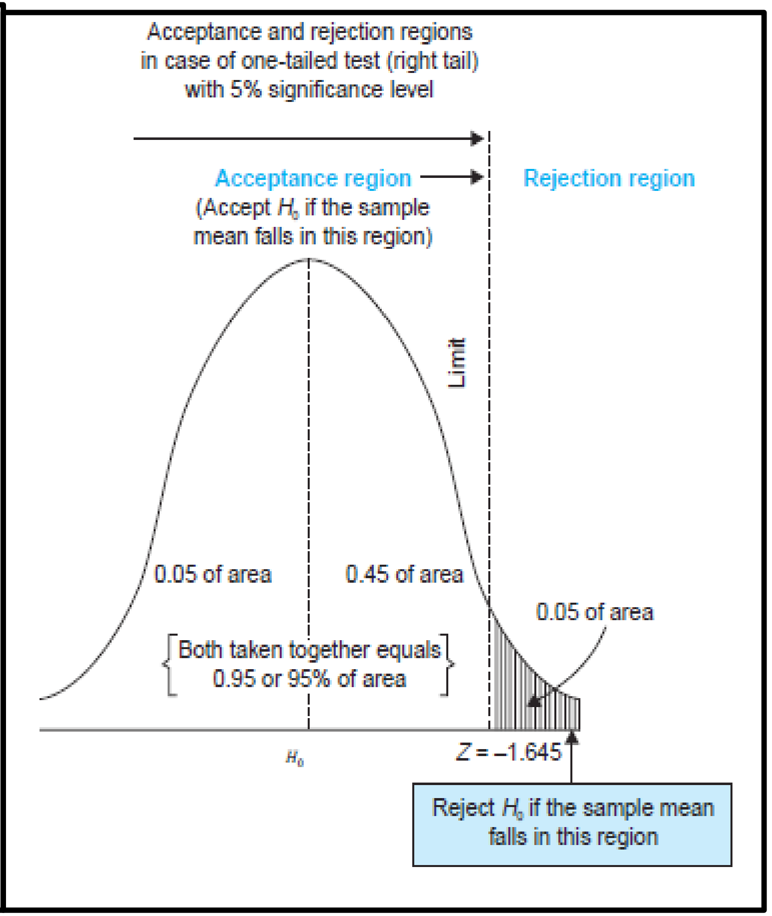
$H_0: \mu = \mu_{H0}$

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



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

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



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

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

Steps in Hypothesis Testing

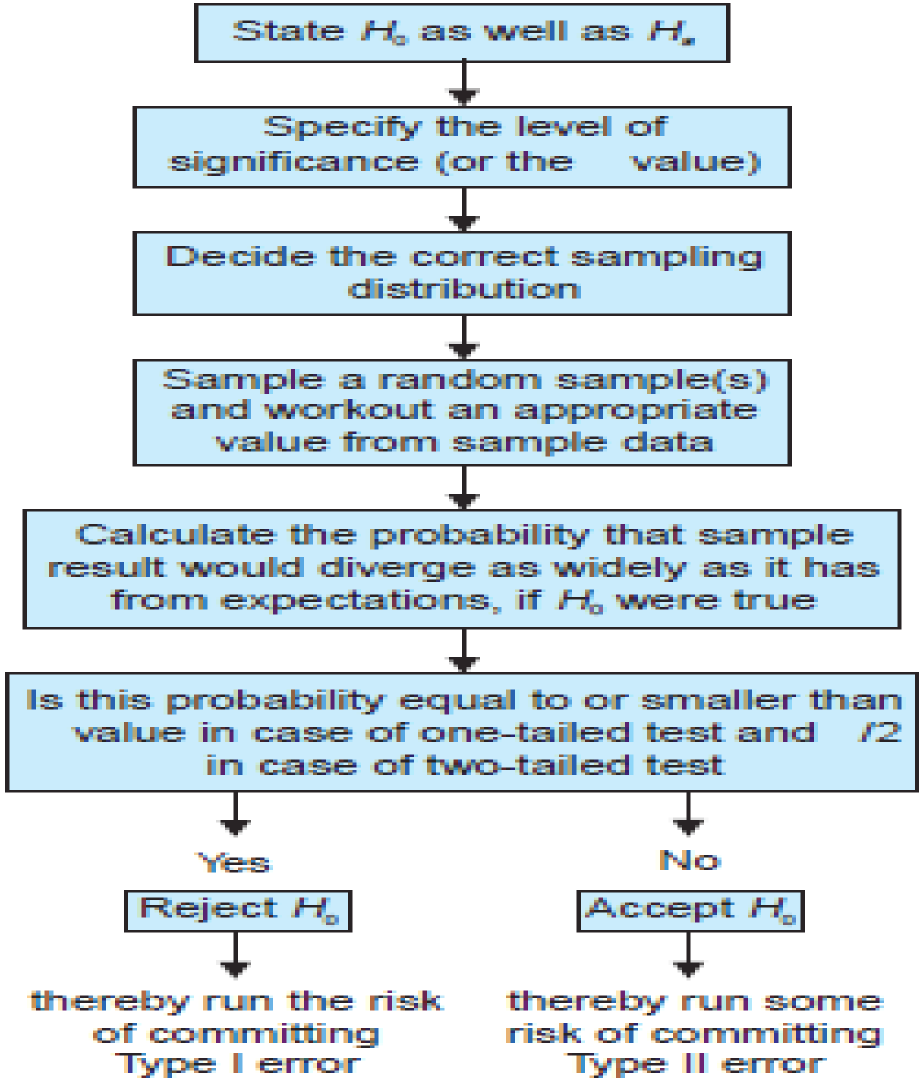
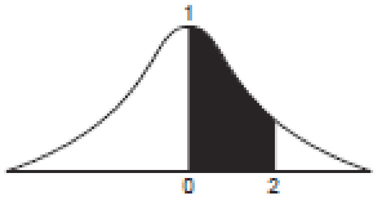


Table 1: Area Under Normal Curve

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.

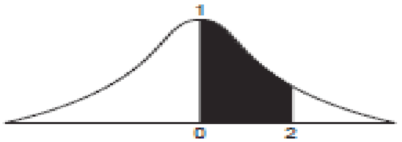


Areas of a standard normal distribution

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

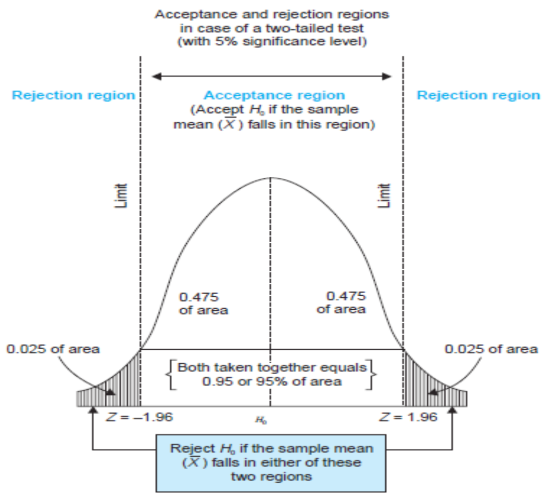
z	.0	0.01	.02	.03	.04	.05	.06	.07	.08	.09
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.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)

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

5. Calculate Probability (P)
6. P < α (one tailed)

   P < α/2 (two tailed)

   Yes => Reject H_0

   (Statistically Significant)

   No => Accept H_a

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

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 $\alpha = 5\%$)

ii) What is the power of the test for $\mu = 105$.

1. State H0 and HA
2. Specify α
3. Choose sampling distribution
4. Calculate test statistic (Z_c)
5. Calculate Probability (P)
6. $P < \alpha$ (one tailed)

   $P < \alpha/2$ (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.

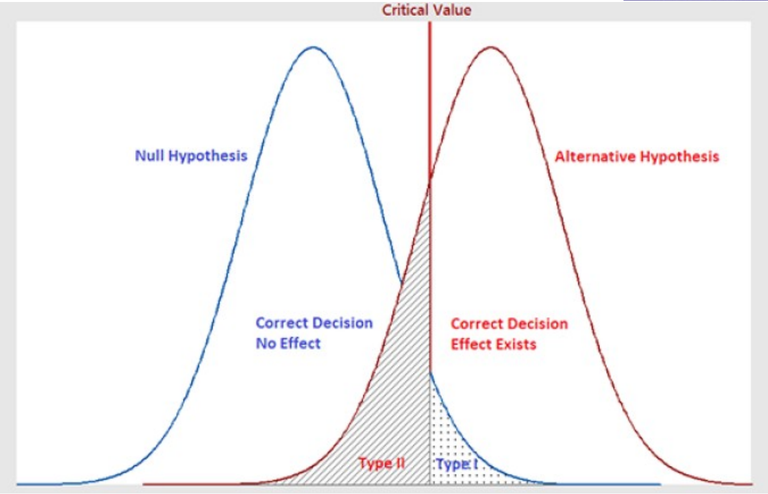
i) Has the wastage increased at a significance level of 10%.
ii) Compute the power of the test for μ = 16.
iii) If the significance level is increased to 20%, what is the new power of the test for μ = 16.?

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.

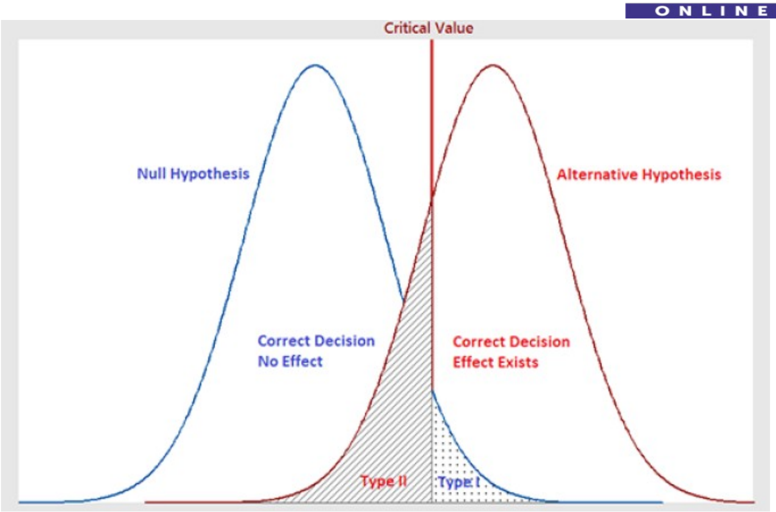
β = Failure to choose Ha when Ha is True. (desirable to be a low value)

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:

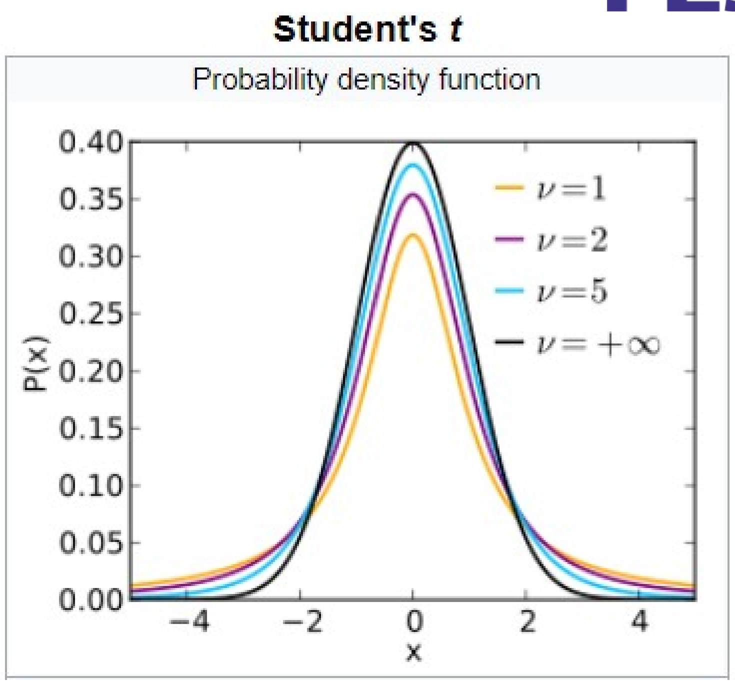
$$z = \frac{\bar{X} - \mu_{H_0}}{\sigma_p / \sqrt{n}}$$

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:

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

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



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).

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

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

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

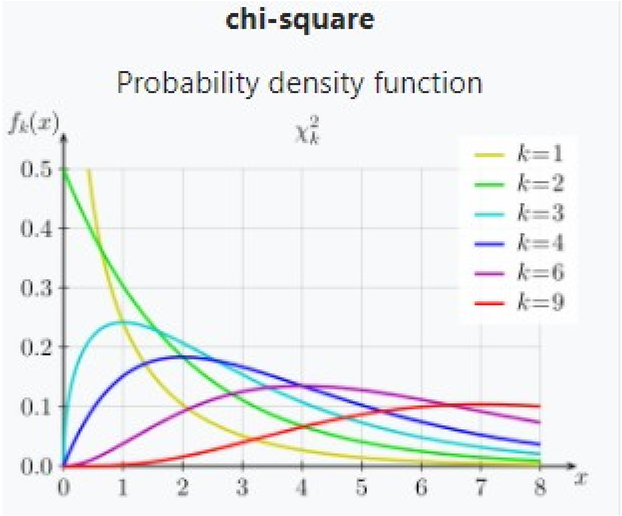
Number of Independent value which are assigned to statistical distribution (n-1) . Eg: Tossing of a die 132 times

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 ith row and jth column

=Expected freqλuency 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)

$P(M) = ¿$

$P(A) = ¿$

$P(A \cap B) = ¿$

$P(M \cap A) = ¿$

$E_{AB} = ¿$

$E_{11} = ¿$

$E_{11} = ¿$ λ

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

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)

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

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

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

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

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

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

	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

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

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

χ^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

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

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

Clas s	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.

[T]he calculated value of χ^2 is much less than this table and hence it is insignificant and can be [a]scribed due to fluctuations of sampling. The result, thus, supports the hypothesis and we may say [t]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 percent-age (Dead)	5.15%	16.36%	10.4%	

H0 : fatality and shirt color are related
Ha : fatality is not related to shirt color.

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

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

@ 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