



PROBABILITY AND STATISTICS

[20A54202]

LECTURER NOTES

I B.TECH & II-SEM

PREPARED BY:

Prof A.PRAKASH

VEMU INSTITUTE OF TECHNOLOGY

[Approved By AICTE, New Delhi and Affiliated to **JNTUA**, Ananthapuramu]

Accredited By NAAC, NBA(EEE, ECE & CSE) & ISO: 9001-2015 Certified Institution

Near Pakala, P.Kothakota, Chittoor- Tirupathi Highway

Chittoor, Andhra Pradesh-517 112

Web Site: **www.vemu.org**



Jawaharlal Nehru Technological University Anantapur
(Established by Govt. of A.P., Act. No. 30 of 2008)
Ananthapuramu–515 002 (A.P) India

Course Code	PROBABILITY AND STATISTICS	L	T	P	C
20A54202		3	0	0	3
Pre-requisite	MATHEMATIS-I	SEMESTER			II

- Course Objectives:
- To familiarize the students with the foundations of probability and statistical methods
- To impart probability concepts and statistical methods in various applications Engineering
- **Unit 1:** Descriptive statistics Statistics Introduction, Measures of Variability (dispersion) Skewness, Kurtosis, correlation, correlation coefficient, rank correlation, principle of least squares, method of least squares, regression lines, regression coefficients and their properties.
- **Learning Outcomes:** At the end of this unit, the student will be able to
 - summarize the basic concepts of data science and its importance in engineering (L2)
 - analyze the data quantitatively or categorically , measure of averages, variability (L4)
 - adopt correlation methods and principle of least squares, regression analysis (L5)
- **UNIT 2:** Probability Probability, probability axioms, addition law and multiplicative law of probability, conditional probability, Baye's theorem, random variables (discrete and continuous), probability density functions, properties.

- ***Learning Outcomes:*** At the end of this unit, the student will be able to
 - Define the terms trial, events, sample space, probability, and laws of probability (L1)
 - Make use of probabilities of events in finite sample spaces from experiments (L3)
 - Apply Baye's theorem to real time problems (L3)
 - Explain the notion of random variable, distribution functions and expected value(L2)
- **UNIT 3:** Probability distributions Discrete distribution - Binomial, Poisson approximation to the binomial distribution and their properties. Continuous distribution: normal distribution and their properties.
Learning Outcomes: At the end of this unit, the student will be able to
 - Apply Binomial and Poisson distributions for real data to compute probabilities, theoretical frequencies
 - Interpret the properties of normal distribution and its applications (L2)
- **Unit 4:** Estimation and Testing of hypothesis, large sample tests Estimation-parameters, statistics, sampling distribution, point estimation, Formulation of null hypothesis, alternative hypothesis, the critical and acceptance regions, level of significance, two types of errors and power of the test. Large Sample Tests: Test for single proportion, difference of proportions, test for single mean and difference of means. Confidence interval for parameters in one sample and two sample problems
- ***Learning Outcomes:*** At the end of this unit, the student will be able to
 - Explain the concept of estimation, interval estimation and confidence intervals (L2)
 - Apply the concept of hypothesis testing for large samples (L4)

- **Unit 5:** Small sample tests Student t-distribution (test for single mean, two means and paired t-test), testing of equality of variances (F-test), χ^2 - test for goodness of fit, χ^2 - test for independence of attributes.
- **Learning Outcomes:** At the end of this unit, the student will be able to
 - Apply the concept of testing hypothesis for small samples to draw the inferences (L3)
 - Estimate the goodness of fit (L5)
- **Reference Books:** 1. S. Ross, a First Course in Probability, Pearson Education India, 2002. 2. W. Feller, an Introduction to Probability Theory and its Applications, 1/e, Wiley, 1968. 3. Peyton Z. Peebles ,Probability, Random Variables & Random Signal Principles -, McGraw Hill Education, 4th Edition, 2001.

UNIT-1 Descriptive statistics

Lecture Notes

1. Descriptive Statistics.

Statistics: Statistics is a tool in the hands of mankind

To translate complex facts into simple and understandable statement of facts.

Statistical Methods:-

1. Collection of Data: The first step of an investigation is the collection of data. Carefull collection is needed because further analysis is based on this.

2. Organisation of Data: The large mass of figures that are collected from a Survey needs organisation.

3. Presentation of Data: The collected data must be edited very carefully so that irrelevant answers and wrong computations must be corrected or adjusted.

The collected data must be classified and tabulated before they can be analysed.

4. Analysis of Data: After presentation of the data the next step is to analyse the presented data. Analysis includes Condensation, Summarisation, Conclusion etc, through means of Measures of Central Tendencies, Dispersion, Skewness, Kurtosis, Correlation and Regression etc.

5. Interpretation of Data: Valid conclusions must be drawn on the basis of analysis. Correct interpretation leads to valid conclusion.

Collection of data : Collection of data is the process of enumeration together with the proper recording of results. Statistical data may be classified as primary and Secondary.

1. Primary data : If an individual or an officer collects the data to study a particular problem, the data are the raw materials of the enquiry. They are the primary data collected by the investigator himself to study any particular problem.

2. Secondary data : Secondary data are those which are already collected by someone for some purpose and are available for the present study.

Ex : The data collected during census operations and are primary data to the department of census and the same data, if used by a research worker for some study are the secondary data.

Sources of Secondary data :

1. published Sources : Such as international publications, official publications of central and state governments, semi-official publications of semi-government institutions like municipal corporations, panchayats etc, publications of research institutions, publications of commercial and financial institutions, reports of various committees, journals and news papers.

2. unpublished Sources: They are records maintained by various government and private offices, the research carried out by individual research scholars in the universities or research institutes.

Population, Sample: In statistical enquiry, all the items, which fall within the preview of enquiry are known as universe or population. That is population is a complete set of all possible observations of the type which is to be investigated. This is a statistical usage and the term population does not necessarily refer to people.

A Sample refers to a smaller, manageable version of a larger group. It is a subset containing the characteristics of a larger population.

Finite and Infinite population: When the number of observations can be counted and definite, it is known as finite population.

Ex: When we are studying the economic background of students of a college, all the students of the college will constitute population and this number will be finite.

When the number of observations cannot be counted and is infinite, it is known as infinite population.

Ex: The number of stars in the sky is infinite population.

Methods of Sampling:

1. Random Sampling Method (Probability Sampling)

A random sample is one where each item in the universe has an equal chance of known opportunity.

2. Non-Random Sampling: This can be done in three methods.

(a). Judgement or purposive Sampling:

The choice of the sampling items depends on the judgement of the investigation.

(b) Quota Sampling: To collect data, the universe is divided into quota according to some characteristics.

(c): Convenience Sampling or check Sampling;

The Sampling is obtained by selecting convenient population units.

1. It is suitable when the population is not clearly defined.

2. Sample is not clear.

3. Complete source list is not available.

Measures of central Tendency and variability.

Measures of Central Tendency: The importance of statistical analysis is to find a number which represents in some definite way the entire data. Such a representative number is called the central value or an average. The value of an average lies somewhere in between the two extreme items possibly in the centre where most of the items concentrate. Hence an average constitutes a measure of the central tendency of the series.

For ungrouped data:

1. Mean or Arithmetic Mean or Expected Value or Average.

$$\therefore \text{Mean } \bar{x} = \frac{\text{Sum of the Scores}}{\text{Number of Scores}} = \frac{\sum x}{n}.$$

→ Find the mean of 40, 45, 48, 57, 78.

Sol: Given data is 40, 45, 48, 57, 78.

$$\text{Mean } \bar{x} = \frac{\sum x}{n} = \frac{40 + 45 + 48 + 57 + 78}{5}$$

$$\therefore \bar{x} = \frac{268}{5} = 53.6.$$

H.W.

→ Find the mean of (i) 45, 55, 50, 45, 40, 55, 45, 50.

$$(ii) 4.5, 2.7, 3.8, 4.9, 3.8, 7.1. \quad 48.125$$

$$4.466$$

2. Median: First write the data in increasing or decreasing order.

If number of observations are even,

$$\text{Median}(\bar{x}) = \frac{\left(\frac{n}{2}\right)^{\text{th}} \text{Term} + \left(\frac{n}{2}+1\right)^{\text{th}} \text{Term}}{2}$$

If number of observations are odd,

$$\text{Median}(\bar{x}) = \left(\frac{n+1}{2}\right)^{\text{th}} \text{Term}$$

→ Find the median of 57, 58, 61, 42, 38, 65, 72, 66.

Sol: Given data is 57, 58, 61, 42, 38, 65, 72, 66.

The data in increasing order is

$$38, 42, 57, 58, 61, 65, 66, 72.$$

In the given data number of observations are even i.e $n=8$.

$$\therefore \text{Median}(\bar{x}) = \frac{\left(\frac{n}{2}\right)^{\text{th}} \text{Term} + \left(\frac{n}{2}+1\right)^{\text{th}} \text{Term}}{2}$$

Put $n=8$

$$\begin{aligned} \therefore \text{Median}(\bar{x}) &= \frac{\left(\frac{8}{2}\right)^{\text{th}} \text{Term} + \left(\frac{8}{2}+1\right)^{\text{th}} \text{Term}}{2} \\ &= \frac{4^{\text{th}} \text{Term} + 5^{\text{th}} \text{Term}}{2} \end{aligned}$$

$$\therefore \text{Median } \bar{x} = \frac{58+61}{2} = \frac{119}{2} = 59.5.$$

→ Find the median of 5, 8, 12, 7, 20, 28, 10.

Sol: Given data is 5, 8, 12, 7, 20, 28, 10.

The data in increasing order is

$$5, 7, 8, 10, 12, 20, 28.$$

If the number of observations are odd i.e $n = 7$,

$$\text{Median}(\bar{x}) = \left(\frac{n+1}{2}\right)^{\text{th}} \text{ term. put } n=7$$

$$\therefore \text{Median}(\bar{x}) = \left(\frac{7+1}{2}\right)^{\text{th}} \text{ term} = \left(\frac{8}{2}\right)^{\text{th}} \text{ term} = 4^{\text{th}} \text{ term.}$$

$$\therefore \text{Median}(\bar{x}) = 10.$$

H.W: → Find the median of (i) : 4, 6, 9, 3, 10, 13, 2, 6

(ii) 22, 24, 30, 27, 29, 31, 25, 28, 41, 42. 28.5

3. Mode: The mode of ungrouped data is most repeated number of observation.

→ Find the mode of 850, 750, 600, 825, 850, 725, 600, 850, 640, 530.

Sol: In the given data, 600 is repeated 2 times.

850 is repeated 3 times.

∴ Most repeated number is 850 ie 3 times.

Hence $\text{Mode}(\bar{x}) = 850$.

→ Find the mode of 41, 45, 48, 57, 78.

Sol: There is no repetition in the given data. Hence
There is no mode for the data.

→ Find the mode of 45, 55, 50, 45, 40, 55, 45, 55.

Sol: In the given data 45 is repeated 3 times and 55 is repeated 3 times.

∴ $\text{Mode(i)} = 45, \text{Mode(ii)} = 55$.

H.W Find the mode of (i) 3, 6, 10, 4, 9, 10. 10

(ii): 6, 7, 10, 12, 13, 12, 48, 12. 12

Grouped Data:

1. Arithmetic mean: $A.M = \frac{\sum f_i x_i}{\sum f_i} = \frac{\sum f x}{N}$

Where $N = \sum f_i$ Total Frequency.

→ From the following, find the mean profits.

Profit per Shop (Rs)	100-200	200-300	300-400	400-500	500-600
Number of shops	10	18	20	26	30
	600-700	700-800			
	28	18			

Sol:

$$\text{Arithmetic Mean} (\bar{x}) = A + \left[\frac{\sum f d}{N} \right]$$

Where, A is Assumed mean or value.

N = Total frequency

d = deviation from the assumed mean.

Calculation Table

Profit per shop (or) Class Interval	Mid Value x	$d = x - A$	Number of shops (or) f	fd
100 - 200	$\frac{100+200}{2} = 150$	-300	10	-3000
200 - 300	$\frac{200+300}{2} = 250$	-200	18	-3600
300 - 400	$\frac{300+400}{2} = 350$	-100	20	-2000
400 - 500	$\frac{400+500}{2} = 450$	0	26	0
500 - 600	$\frac{500+600}{2} = 550$	100	30	3000
600 - 700	$\frac{600+700}{2} = 650$	200	28	5600
700 - 800	$\frac{700+800}{2} = 750$	300	18	5400
			$\sum f = 150$	$\sum fd = 5400$

From the table $A = 450$ (Assumed Mean)

$$\sum f = N = 150.$$

$$\sum fd = 5400.$$

$$\text{Arithmetic Mean } \bar{x} = A + \left[\frac{\sum fd}{N} \right]$$

$$\therefore \bar{x} = 450 + \left[\frac{5400}{150} \right]$$

$$= 450 + 36 = 486. R_s.$$

Average profit is $R_s 486$.

Q Median: Calculate the median from the following data.

Marks	10-25	25-40	40-55	55-70	70-85	85-100
Frequency	6	20	44	26	3	1

Sol: $\text{Median}(M) = l + \left[\frac{\frac{N}{2} - m}{f} \right] \times c$

Table

Marks	Frequency	cumulative Frequency
10-25	6	6
25-40	20	26
40-55	44	70
55-70	26	96
70-85	3	99
85-100	1	100
		$N=100$

First we find $\frac{N}{2} = \frac{100}{2} = 50$.

∴ Here just greater than cumulative value by $\frac{N}{2} = 50$ is 70.

∴ There for 40-55 is Median class.

$l = \text{lower end of the Median class} = 40$

$m = \text{Before class cumulative frequency} = 26$

$c = \text{length of the class interval} = 55-40 = 15$

$f = \text{Same class frequency} = 44$

$$\therefore \text{Median} = 40 + \left[\frac{50-26}{44} \right] \times 15$$

$$= 40 + 8.18 = 48.18$$

→ 3. Note: Find the mode of the following distribution:

class Interval	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
Frequency	5	8	7	12	28	20	10	10

Sol: Mode = $l + \left[\frac{f - f_1}{2f - f_1 - f_2} \right] \times c$

In the given data maximum frequency is 28 i.e 40-50 class is the modal class.

class	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
f	5	8	7	12	28	20	10	10

Where l = lower end of the modal class = 40.

c = length of the class interval = $50 - 40 = 10$

f = frequency of the modal class = 28

f_1 = before class frequency = 12

f_2 = frequency of the class succeeding the modal class = 20

∴ Mode = $40 + \left[\frac{28 - 12}{2 \times 28 - 12 - 20} \right] \times 10 = 40 + 6.66 = 46.66$.

H.W: → Find Median & Mode of the following:

(1) class	10-20	20-30	30-40	40-50	50-60	60-70	70-80
Freq	2	3	8	14	8	3	2

(2)	20-30	30-40	40-50	50-60	60-70	70-80	80-90
Freq	3	61	132	153	140	51	2

Median = 56.90, Mode = 56.17

Measures of Dispersion:

Introduction: If the items within a distribution differ from one another in magnitude the dispersion or scatteredness is used to indicate the difference. The distribution differ from one another in respect of two main characteristics.

1. They may differ in measures of central tendency.
2. They may have the same measure of central tendency but have wide disparities in the formality of distribution.

→ Consider the two Series 3, 4, 5, 6, 7 and 12, 13, 14, 15, 16.

The mean of the two Series are 5 and 14.

Although the means are different the items in the two series are scattered in the same way around the means.

→ Consider the two Series 5, 8, 10, 4, 3 and 6, 15, 0, 7, 2.

The mean of the two Series ~~are~~ is 6 but the scatteredness of the various items in the two series about their mean is different.

From the two examples we infer that the mean fails to give us an idea how the various items are scattered and how the distribution are constituted. Hence we need the Measures of dispersion.

1. Range

2. Mean Deviation

3. Standard Deviation

Ungrouped Data :

1. Range : Range for ~~an~~ ungrouped data is defined as the difference between the greatest and least values of the variate.

$$\therefore \text{Range} = \text{Greatest value} - \text{Smallest value}$$

→ Find the range of marks of students in a class given as

60, 72, 96, 28, 35, 10, 40, 9, 85, 25.

Sol: Range = Largest value - Smallest value = 96 - 9 = 87.

→ The following table gives the daily sales (R_s) of two firms A and B for five days

	Firm A	5050	5025	4950	4835	5140	\bar{x}_A
--	--------	------	------	------	------	------	-------------

	Firm B	4900	3100	2200	1800	13000	\bar{x}_B
--	--------	------	------	------	------	-------	-------------

Sol: The sales of both the firms in average is same but distribution pattern is not similar. Therefore there is a great amount of variation in the daily sales of the firm B than that of the firm A.

$$\begin{aligned}\text{Range of Sales of firm A} &= \text{Greatest value} - \text{Least value} \\ &= 5140 - 4835 = 305\end{aligned}$$

$$\begin{aligned}\text{Range of Sales of firm B} &= \text{Greatest value} - \text{Least value} \\ &= 13000 - 1800 = 11200\end{aligned}$$

2. Mean Deviation: Mean deviation is defined as Arithmetic mean of absolute values of the deviations of the variate measured from an average (Median, Mode, Mean).

The absolute value of the deviation denoted by |deviation| is the numerical value of the deviation with positive sign.

Let x_1, x_2, \dots, x_n be the values of n variables and \bar{x} be their mean. Let $|x_i - \bar{x}|$ be the absolute value of the deviation of the variable x_i from \bar{x} .

$$\therefore \text{Mean Deviation (M.D)} = \frac{\sum_{i=1}^n |x_i - \bar{x}|}{n}.$$

→ Find the mean deviation of the variables 40, 62, 54, 68, 76 from Arithmetic mean.

Sol: Given that 40, 62, 54, 68, 76, $n=5$.

$$\text{Arithmetic Mean } \bar{x} = \frac{\sum x_i}{n} = \frac{40+62+54+68+76}{5} = 60.$$

$$\begin{aligned} \text{M.D.} &= \frac{\sum_{i=1}^n |x_i - \bar{x}|}{n} \\ &= \frac{|40-60| + |62-60| + |54-60| + |68-60| + |76-60|}{5} \end{aligned}$$

$$= \frac{20+2+6+8+16}{5} = \frac{52}{5} = 10.4.$$

3. Standard Deviation (S.D): If we take the mean of the squared deviations from the mean i.e. $\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$, then it is found that this number leads to a proper measure of dispersion. The number is called Variance and is denoted by σ^2 . Then σ the Standard Deviation is given by the positive square root of variance.

$$\therefore \text{Variance} (\sigma^2) = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$$

$$\text{Standard Deviation} (\sigma) = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}}$$

Thus if x_1, x_2, \dots, x_n are n observations and \bar{x} is their mean, we consider $\sum_{i=1}^n (x_i - \bar{x})^2$

Case 1: If $\sum_{i=1}^n (x_i - \bar{x})^2 = 0$, then each $(x_i - \bar{x}) = 0$, which implies that all observations are equal to the mean \bar{x} and there is no dispersion.

Case 2: If $\sum (x_i - \bar{x})^2$ is small, then it shows that each observation x_i is very close to the mean \bar{x} and hence the degree of dispersion is very low.

Case 3: If $\sum (x_i - \bar{x})^2$ is large, then it indicates that a higher degree of dispersion of observations from the mean \bar{x} .

→ Find the variance and standard deviation of the following data 5, 12, 3, 18, 6, 8, 2, 10.

Sol: Given data is 5, 12, 3, 18, 6, 8, 2, 10, $n = 8$

$$\text{Mean } \bar{x} = \frac{\sum x}{n} = \frac{5+12+3+18+6+8+2+10}{8}$$

$$\therefore \bar{x} = 8$$

x	5	12	3	18	6	8	2	10
$x - \bar{x}$	-3	4	-5	10	-2	0	-6	2
$(x - \bar{x})^2$	9	16	25	100	4	0	36	4

$$\therefore \text{Variance } \sigma^2 = \frac{\sum (x - \bar{x})^2}{n} = \frac{194}{8} = 24.25$$

$$\text{Standard Deviation } \sigma = \sqrt{24.25} = 4.95 \text{ (Nearly)}$$

Grouped Data:

(i) Mean Deviation for Discrete Frequency Distribution

$$\text{Mean Deviation (M.D)} = \frac{\sum_{i=1}^n f_i |x_i - \bar{x}|}{N}$$

Where \bar{x} is Mean or Median or Mode

N = Total frequency.

f_i = Frequency of the respective class.

→ Find the Mean Deviation about the mean for the following data

x_i	2	5	7	8	10	35
f_i	6	8	10	6	8	2

Sol: Mean Deviation (M.D) = $\frac{\sum f_i |x_i - \bar{x}|}{N}$

$$\text{Mean}(\bar{x}) = \frac{\sum f_i x_i}{N}$$

x_i	f_i	$f_i x_i$	$ x_i - \bar{x} $	$f_i x_i - \bar{x} $	
2	6	12	6	36	
5	8	40	3	24	
7	10	70	1	10	
8	6	48	0	0	
10	8	80	2	16	
35	2	70	27	54	
$\Sigma f_i = 40$			$\Sigma f_i x_i - \bar{x} = 140$		

$$\text{Mean } (\bar{x}) = \frac{320}{40} = 8$$

$$\text{Mean Deviation (M.D)} = \frac{140}{40} = 3.5$$

→ Find the mean deviation from the median for the following data

x_i	6	9	3	12	15	13	21	22
f_i	4	5	3	2	5	4	4	3

Sol: First we write the data ascending order to get the table as follows

x_i	3	6	9	12	13	15	21	22
f_i	3	4	5	2	4	5	4	3

Here $n = 8$, Median (\bar{x}) = $\frac{(\frac{n}{2})^{\text{th}} + (\frac{n}{2}+1)^{\text{th}} \text{ term}}{2}$

$$\therefore \bar{x} = \frac{(\frac{8}{2})^{\text{th}} + (\frac{8}{2}+1)^{\text{th}} \text{ term}}{2} \text{ msg } x = \frac{4^{\text{th}} + 5^{\text{th}} \text{ term}}{2}$$

$$\therefore \bar{x} = \frac{12+13}{2} = 12.5 \sim 13 \quad (\text{Rounded})$$

x_i	$ x - \bar{x} $	f_i	$f_i x - \bar{x} $
3	10	3	30
6	7	4	28
9	4	5	20
12	1	2	2
13	0	4	0
15	2	5	10
21	8	4	32
22	9	3	27
$N = 30$		$\sum f_i x - \bar{x} = 149$	

$$M.D = \frac{\sum f_i |x_i - \bar{x}|}{N} = \frac{149}{30} = 4.97.$$

→ Find the Mean Deviation from the Mean by Step Deviation

(Q) Short cut Method.

Classes	0-100	100-200	200-300	300-400	400-500	500-600	600-700	700-800
Freq	4	8	9	10	7	5	4	3

Sol: Mean deviation ($M.D$) = $\frac{\sum f|x - \bar{x}|}{N}$

Where $\bar{x} = \text{Mean by Shortcut Method} = A + \left[\frac{\sum fd}{N} \right] \times c$

Where $d = \frac{x - \text{Assumed mean}}{\text{class size i.e } c} = \frac{x - A}{c}$

class	Midvalue x	$d = \frac{x - A}{c}$	f	fd	$ x - \bar{x} $	$f x - \bar{x} $
0-100	50	-3	4	-12	308	1232
100-200	150	-2	8	-16	208	1664
200-300	250	-1	9	-9	108	972
300-400	350 ^(A)	0	10	0	8	80
400-500	450	1	7	7	92	644
500-600	550	2	5	10	192	960
600-700	650	3	4	12	292	1168
700-800	750	4	3	12	392	1176

Where $d = \frac{x_i - 350}{100}$

$N = 50 \quad \sum fd = 4 \quad \sum f|x - \bar{x}| = 7896$

$C = 100 - 0 = 100$
 $400 - 300 = 100$

$\bar{x} = 350 + \left[\frac{4}{50} \right] \times 100 = 358$

$M.D = \frac{7896}{50} = 157.92$

(ii) Continuous Frequency Distribution:

Method 1 Variance $\sigma^2 = \frac{\sum f_i(x_i - \bar{x})^2}{N}$

$$\text{Standard Deviation } \sigma = \sqrt{\frac{\sum f_i(x_i - \bar{x})^2}{N}}$$

Method 2. Variance $\sigma^2 = \frac{\sum f_i x_i^2}{N} - \left[\frac{\sum f_i x_i}{N} \right]^2$

$$\text{Standard Deviation } \sigma = \sqrt{\frac{\sum f_i x_i^2}{N} - \left[\frac{\sum f_i x_i}{N} \right]^2}$$

Method 3: Standard Deviation Method or Shortcut Method:

$$\text{Variance } \sigma^2 = \frac{h^2}{N^2} \left[N \sum f_i d_i^2 - (\sum f_i d_i)^2 \right]$$

$$\text{Standard Deviation } \sigma = \sqrt{\frac{h^2}{N} \left[N \sum f_i d_i^2 - (\sum f_i d_i)^2 \right]}$$

→ Calculate the variance and S.D of the following distribution.

Classes :	30-40	40-50	50-60	60-70	70-80	80-90	90-100
Freq :	3	7	12	15	8	3	2

Sol:

$$\text{Variance } \sigma^2 = \frac{c^2}{N^2} \left[N \sum f_i d_i^2 - (\sum f_i d_i)^2 \right]$$

$$\text{∴ Standard Deviation } \sigma = \frac{c}{N} \sqrt{N \sum f_i d_i^2 - (\sum f_i d_i)^2}$$

$$\text{Mean } \bar{x} = A + \left[\frac{\sum f_i d_i}{N} \right] \times c$$

Where $A = \text{Assumed Mean}$.

∴ $\sigma = \text{Standard deviation of original data}$

class	Freq: f_i	Mid value x_i	$d_i = \frac{x_i - A}{C}$ $A=65, C=10$	d_i^2	$f_i d_i$	$f_i d_i^2$
30-40	3	35	-3	9	-9	27
40-50	7	45	-2	4	-14	28
50-60	12	55	-1	1	-12	12
60-70	15	65	0	0	0	0
70-80	8	75	1	1	8	8
80-90	3	85	2	4	6	12
90-100	2	95	3	9	6	18
$N = 50$				$\sum f_i d_i = -15$	$\sum f_i d_i^2 = 105$	

Assumed mean $A = 65$

Length of the class Interval $C = 70-60 = 10$

$$\text{Mean } \bar{x} = A + \left[\frac{\sum f_i d_i}{N} \right] \times C$$

$$= 65 + \left[\frac{-15}{50} \right] \times 10 = 65 - 3 = 62$$

$$\text{Variance } \sigma^2 = \frac{C^2}{N^2} \left[N \sum f_i d_i^2 - (\sum f_i d_i)^2 \right]$$

$$= \frac{10^2}{50^2} \left[50(-15) - (-15)^2 \right]$$

$$= \frac{100}{2500} [5250 - 225] = 201$$

$$\text{Standard Deviation } \sigma = \sqrt{201} = 14.18$$

Coefficient of variance:

Coefficient of variation (C.V) is defined as the ratio of the standard deviation σ to the arithmetic mean \bar{x} and it is often expressed as a percentage.

$$\therefore \text{Coefficient of Variance} = \frac{\sigma}{\bar{x}} \times 100 \quad \text{where } \bar{x} \neq 0$$

Comparison:

Suppose two distributions are having same mean $\bar{x}_1 = \bar{x}_2 = \bar{x}$ but different standard deviations σ_1 and σ_2 respectively. Then coefficient of variance are given by $(\frac{\sigma_1}{\bar{x}} \times 100)$ and $(\frac{\sigma_2}{\bar{x}} \times 100)$. Thus the C.V's can be compared using σ_1 and σ_2 only.

Here, the series with lower value of standard deviation is said to be more consistent than the other series with greater standard deviation. The series with greater standard deviation is called more dispersed than other.

- The scores of two cricketers A and B in 10 innings are given here. Find who is better run getter and who is more consistent player.

Scores of A x_i	40	25	19	80	38	8	67	121	66	76
Scores of B y_i	28	70	31	0	14	111	66	31	25	4

Sol: Standard Deviation $A = \sigma_x = \sqrt{\frac{\sum (x-\bar{x})^2}{n}}$, $\sigma_B = \sqrt{\frac{\sum (y-\bar{y})^2}{n}}$

Coefficient of Variance of A = $\frac{\sigma_x}{\bar{x}} \times 100$, C.V of B = $\frac{\sigma_y}{\bar{y}} \times 100$.

$$\bar{x} = \frac{\sum x}{n}, \bar{y} = \frac{\sum y}{n}$$

x	(x - \bar{x})	$(x - \bar{x})^2$	y _i	y - \bar{y}	$(y - \bar{y})^2$
40	-14	196	28	-10	100
25	-29	841	70	32	1024
19	-35	1225	31	-7	49
80	26	676	0	-38	1444
38	-16	256	14	-24	576
8	-46	2116	111	73	5329
67	13	169	66	28	784
121	67	4489	31	-7	49
66	12	144	25	-13	169
76	2	484	4	-34	1156
$\sum x = 540$		$\sum (x - \bar{x})^2 = 10596$	$\sum y = 380$		$\sum (y - \bar{y})^2 = 10,674$

For Cricketer A, $\bar{x} = \frac{540}{10} = 54$, $\bar{y} = \frac{380}{10} = 38$.

Standard Deviation of

$$A, \sigma_x = \sqrt{\frac{10596}{10}} = 32.55$$

$$B, \sigma_y = \sqrt{\frac{10,674}{10}} = 32.67$$

$$\text{C.V of } A = \frac{\sigma_x}{\bar{x}} \times 100 = \frac{32.55}{54} \times 100 = 60.28$$

$$\text{C.V of } B = \frac{32.67}{38} \times 100 = 86.68$$

Since $\bar{x} > \bar{y}$, Cricketer A is a better cricketer.

But C.V of A < C.V of B.

∴ Cricketer A is more consistent player.

Skewness: Def:

A distribution which is not symmetrical is called a skewed distribution. In such distribution the mean, the mode and median will not coincide. The values are pulled apart.

Explain the Skewness: The measures of central tendency and dispersion do not indicate whether the distribution is symmetric or not. Measures of Skewness gives the direction and the extent of skewness. In symmetrical distribution the mean, median and mode are identical. Thus Skewness is the lack of symmetry. The measures of central tendency and dispersion are inadequate to characterise a distribution completely. They may be supported by two more measures Skewness and Kurtosis.

Test of Skewness: If the distribution is symmetric the following conditions are observed.

1. The values of Mean, Median, Mode coincide (Equal)
2. $Q_3 - \text{Median} = \text{Median} - Q_1$
3. The sum of positive deviations = The sum of negative deviations.
4. The frequencies on either side of the mode are equal

Similarly a Skewed distribution will have the following characteristics:

1. Mean \neq Median \neq Mode.

2. $Q_3 - \text{Median} \neq \text{Median} - Q_1$

3. The sum of positive deviation \neq The sum of negative deviations.

Measures of Skewness:

Absolute Skewness = Mean - Mode

If the value of Mean is greater than mode then the Skewness is positive.

If the value of Mode is greater than Mean the Skewness is negative.

There are three important measures of Relative Skewness

1. Karl Pearson's Coefficient of Skewness.

$$Sk_p = \frac{\bar{x} - \text{Mode}}{\sigma}, \quad Sk_p = \frac{3(\text{Mean} - \text{Median})}{\sigma}$$

2. Bowley's Coefficient of Skewness.

3. Kelly's Coefficient of Skewness.

→ Calculate Karl Pearson's Coefficient of Skewness for the following data 25, 15, 23, 40, 27, 25, 23, 25, 20.

Sol: Karl Pearson Coefficient of Skewness = $\frac{\text{Mean} - \text{Mode}}{\text{S.D}}$

Mean = $\frac{\sum x}{n}$, Mode = Most repeated Value.

$$\text{S.D} = \sqrt{\frac{\sum (x-\bar{x})^2}{n}}$$

x	$x - \bar{x}$	$(x - \bar{x})^2$
25	0	0
15	-10	100
23	-2	4
40	15	225
27	2	4
25	0	0
23	-2	4
25	0	0
20	-5	25
$\Sigma x = 223$		$\Sigma (x - \bar{x})^2 = 362$

$$\bar{x} = \frac{223}{9} = 24.77 \approx 25, \quad \text{Mode} = 25 \quad (\text{repeated 3 times})$$

$$\text{S.D} = \sqrt{\frac{362}{9}} = 6.3$$

$$S_{kp} = \frac{24.77 - 25}{6.3} = -\frac{0.23}{6.3} = -0.03$$

Correlation:

Def: Correlation is a statistical analysis which measures and analysis the degree or extent to which two variables fluctuate with reference to each other
(OR)

Correlation refers to the relationship of two or more variable

Ex: Relationship between the height of a father and a son, wage and price index.

Types of Correlation: Correlation is classified into many types,

1. positive and negative Correlation
2. Simple and Multiple Correlation
3. partial and Total Correlation
4. Linear and Non-Linear Correlation.

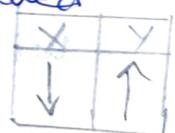
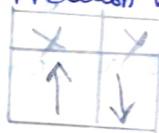
1. positive and negative Correlation;

→ Positive Correlation: If two variables tend to move together in the same direction ie an increase (decrease) in the value of one variable is accompanied by an increase (decrease) in the value of the other variables, then the Correlation is called positive or direct Correlation.

X	Y
↑	↑

X	Y
↓	↓

→ Negative Correlation: If two variables tend to move together in opposite directions so that an increase or decrease in the values of one variable is accompanied by a decrease or increase in the value of the other variable, then the correlation is called Negative or Inverse Correlation.



2. Simple and Multiple Correlation:

When we study only two variables, the relation is described as Simple Correlation.

Ex: Quantity of money and price level, demand and price etc.
We study more than two variables simultaneously called Multiple Correlation.

Ex: The relationship of Price, demand and Supply of a commodity.

3. Partial and Total Correlation:

The study of two variables excluding some other variables is called Partial Correlation.

Ex: We study Price and Demand eliminating Supply. We study the total Correlation if all the factors are taken into account.

4. Linear and Non-Linear Correlation:

If the ratio of change between two variables is uniform then there will be linear Correlation between them.

Ex: We can see that the ratio of change b/w the variables is the same. If we plot these on the graph, we get a straight line.

A 2 7 12 17
B 3 9 15 21

In a Curvilinear or Non-Linear Correlation the amount of change in one variable does not bear a constant ratio of the amount of change in the other variable. The graph of non-linear or curvilinear relationship will be a curve.

Methods of Studying Correlation:

There are two different methods for finding out the relationship between variables. They

are (1) Graphic Method (2) Mathematical Method.

1. Graphic Methods are:

(a) Scatter diagram or Scattergraph.

(b) Simple graph.

2. Mathematical Methods are:

(a) Karl Pearson's Coefficient of Correlation.

(b) Spearman's Rank Correlation Coefficient.

(c) Coefficient of Concurrent deviation.

(d) Method of least Squares.

Covariance: Let X and Y be the two random variables with means \bar{X} and \bar{Y} respectively. Then the Covariance between two variables X and Y is defined by the relation $\text{Cov}(X, Y) = E[(X - \bar{X})(Y - \bar{Y})]$.

$$= E(XY) - E(X) \cdot E(Y).$$

If x and y are two Independent random variables

then $\text{Cov}(x, y) = 0$.

If $\text{Cov}(x, y) \neq 0$ then the two variables x and y are Dependent or Non-Independent.

Note: $\text{Cov}(x+a, y+b) = \text{Cov}(x, y)$.

$\text{Cov}(ax, by) = ab \text{Cov}(x, y)$.

$\text{Cov}(x, y) = \frac{1}{n} \sum x_i y_i - \bar{x} \bar{y}$.

Karl Pearson's coefficient of Correlation or Product-Moment Correlation Coefficient:

Karl Pearson's coefficient of Correlation γ between the Variables x and y is defined by the formula

$$\gamma = \frac{\text{Cov}(x, y)}{(S.D. of x)(S.D. of y)}$$

$$\gamma = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n \sigma_x \sigma_y}$$

$$\text{where } \sigma_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$$

OR

$$\gamma = \frac{\frac{1}{n} \sum x_i y_i - \bar{x} \bar{y}}{\sqrt{\frac{\sum x_i^2}{n} - (\bar{x})^2} \sqrt{\frac{\sum y_i^2}{n} - (\bar{y})^2}}$$

$$\bar{x} = \frac{\sum x}{n}$$

$$\bar{y} = \frac{\sum y}{n}$$

$$\gamma = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \sqrt{n \sum y_i^2 - (\sum y_i)^2}}$$

Properties of Correlation Coefficient:

1. The coefficient of correlation lies between -1 and 1 i.e. $-1 \leq r \leq 1$ or $|r| \leq 1$.
2. If $r = 1$, Correlation is perfect and positive.
3. If $r = -1$, Correlation is perfect and negative.
4. If $r = 0$, then there is no relation between the variables.
5. The coefficient of correlation is independent of the change of origin and scale of measurements.
6. $r(ax+b, cy+d) = \frac{ac}{\sqrt{ac}} r(x, y)$.
7. Two independent variables are uncorrelated. That is x and y are independent variables $r(x, y) = 0$.

→ Find the Coefficient of Correlation between the two.

Variables	x	50	50	55	60	65	65	60	60	60
	y	11	13	14	16	16	15	15	14	13

Sol: Correlation coefficient $r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\sum x^2 - \frac{(\sum x)^2}{n}} \sqrt{\sum y^2 - \frac{(\sum y)^2}{n}}}$

x	y	$x = x - \bar{x}$	$y = y - \bar{y}$	x^2	y^2	xy
50	11	-9	-3	81	9	27
50	13	-9	-1	81	1	9
55	14	-4	0	16	0	0
60	16	1	2	1	4	2
65	16	6	2	36	4	12
65	15	6	1	36	9	15
65	15	6	1	36	1	6
60	14	1	0	1	0	0
60	13	1	-1	1	1	-1
60	13	1	-1	1	1	-1
590	140	0	0	290	22	60
Σx	Σy	Σx	Σy	Σx^2	Σy^2	Σxy

$$\bar{x} = \frac{\sum x}{n} = \frac{590}{10} = 59, \quad \bar{y} = \frac{\sum y}{n} = \frac{140}{10} = 14.$$

$$r = \frac{60 - 0}{\sqrt{290 - 0} \sqrt{22 - 0}} = 0.75.$$

∴ Between x and y high positive Correlation.

Rank Correlation Coefficient:

Spearman's rank correlation coefficient is given by

$$P = 1 - \left[\frac{6 \sum d^2}{n(n^2-1)} \right]$$

Where P is Spearman's rank correlation coefficient.

d^2 = Sum of the Squares of the differences of two ranks.

n = number of paired observations.

Properties of rank Correlation Coefficients:

1. The value of P lies between -1 and $+1 \Rightarrow -1 \leq P \leq 1$.

2. If $P = 1$, there is Complete agreement in the order if the ranks and the direction of the rank is Same.

3. If $P = -1$, then there is a Complete disagreement in the order of the ranks and they are in opposite direction.

→ The following are the ranks obtained by 10 students in two Subjects, Statistics and Mathematics. To what extent the knowledge of the students in two subjects is related?

Statistics	1	2	3	4	5	6	7	8	9	10
Mathematics	2	4	1	5	3	9	7	10	6	8

Sol: Rank Correlation Coefficient $P = 1 - \left[\frac{6 \sum d^2}{n(n^2-1)} \right]$

x	1	2	3	4	5	6	7	8	9	10
y	2	4	1	5	3	9	7	10	6	8
$d = x - y$	-1	-2	2	-1	2	-3	0	-2	3	2
d^2	1	4	4	1	4	9	0	4	9	4
$\sum d^2 = 40$										

$$\therefore P = 1 - \left[\frac{6 \times 40}{10(10^2-1)} \right] = 1 - \left[\frac{240}{990} \right] = 0.76.$$

Comment: The knowledge between Statistics and Mathematics are high positive correlation.

→ Ten competitors in a musical test were ranked by the three judges A, B and C in the following order,

Ranks by A	1	6	5	10	3	2	4	9	7	8
Ranks by B	3	5	8	4	7	10	2	1	6	9
Ranks by C	6	4	9	8	1	2	3	10	5	7

Using the rank Correlation method, discuss which pair of judges has the nearest approach to common liking in music.

$$\text{Sol: Rank Correlation Coefficient } P = 1 - \left[\frac{6 \sum d^2}{n(n^2-1)} \right]$$

Where $n = 10$

Let Ranks by A = x , Ranks by B = y , Ranks by C = z

x	y	z	$d_1 = x - y$	$d_2 = \frac{x-z}{y-z}$	$d_3 = \frac{y-z}{x-z}$	d_1^2	d_2^2	d_3^2
1	3	6	-2	-5	-3	4	25	9
6	5	4	1	2	1	1	4	1
5	8	9	-3	-4	-1	9	16	1
10	4	8	6	2	-4	36	4	1
3	7	1	-4	2	6	16	4	1
2	10	2	-8	0	8	64	0	36
4	2	3	2	1	-1	4	1	1
9	1	10	-8	-1	-9	64	1	81
7	6	5	1	2	1	4	1	4
8	9	7	-1	1	2	1	1	4

$$P_1(x,y) = 1 - \left[\frac{6 \times 200}{10(10^2-1)} \right] = 1 - \frac{40}{33} = -\frac{7}{33}$$

$$P_2(x,z) = 1 - \left[\frac{6 \times 60}{10(10^2-1)} \right] = 1 - \frac{4}{11} = \frac{7}{11}$$

$$P_3(y,z) = 1 - \left[\frac{6 \times 214}{10(10^2-1)} \right] = 1 - \frac{214}{165} = -\frac{49}{165}$$

Since $P_2(x,z)$ is maximum, we conclude that the pair of Judges A and C has the nearest approach to common liking.

Regression.

Brief: The statistical method which helps us to estimate the unknown value of one variable from the known value of the related variable is called "Regression". The line described in the average relationship between two variables is known as line of regression or estimating line.

- Uses:
1. It is used to estimate the relation between two economic variables like income and expenditure.
 2. It is a highly valuable tool in Economics and Business.
 3. It is widely used for prediction purpose.
 4. We can calculate coefficient of Correlation and Coefficient of determination with the help of the regression coefficient.
 5. It is useful in statistical estimation of demand curves, Supply curves, production function, cost function and consumption function etc.

Comparison between the Correlation and Regression:

The Correlation Coefficient is a measure of degree of Covariability between two variables, while the regression establishes a functional relation between dependent and independent variables so that the former can be predicted for a given value of the later. In Correlation both the variables x and y are random variables, whereas in regression, x is a random variable and y is fixed variable. The

Correlation coefficient is a relative measure whereas regression coefficient is an absolute figure.

Methods of studying Regression:

1: Graphical Method 2. Algebraic Method

Regression Line by Least-Squares Curve fitting procedure,

Let the set of data points be $(x_i, y_i), i = 1, 2, 3, \dots, n$.

Suppose the curve $y = f(x)$ is fitted to this data.

Let the observed value at $x = x_i$ is y_i and the corresponding value on the curve is $f(x_i)$. Let e_i be the error of approximation at $x = x_i$. Then we have $e_i = y_i - f(x_i) \rightarrow ①$

$$\text{Consider } S = [y_1 - f(x_1)]^2 + [y_2 - f(x_2)]^2 + \dots + [y_n - f(x_n)]^2 \\ = e_1^2 + e_2^2 + \dots + e_n^2. \rightarrow ②$$

The method of least squares consists of minimising S .

Fitting of a straight line:

Let $y = a + bx$ be the straight line to be fitted for the given data. Then $S = \sum [y_i - a - bx_i]^2 \rightarrow ③$

If S is minimum, we have $\frac{\partial S}{\partial a} = 0, \frac{\partial S}{\partial b} = 0$

$$\frac{\partial S}{\partial a} = 0 \Rightarrow 2 \sum [y_i - a - bx_i] (-1) = 0 \quad \sum_{i=1}^n a = a + a + \dots + a \\ \Rightarrow \sum [y_i - a - bx_i] = 0$$

$$\Rightarrow \sum y_i - \sum a - b \sum x_i = 0$$

$$\Rightarrow \sum y_i = n a + b \sum x_i$$

$$\therefore n a + b \sum x_i = \sum y_i \rightarrow ④$$

$$\sum_{i=1}^n a = a + a + \dots + a \\ = na$$

$$\begin{aligned}\frac{\partial S}{\partial b} &= 0 \Rightarrow 2 \sum [y_i - a - bx_i]x_i = 0 \\ &\Rightarrow \sum [y_i - a - bx_i]x_i = 0 \\ &\Rightarrow \sum x_i y_i - a \sum x_i - b \sum x_i^2 = 0 \\ &\Rightarrow a \sum x_i + b \sum x_i^2 = \sum x_i y_i \rightarrow (5)\end{aligned}$$

By Solving (4) + (5) we get a and b. These (4) + (5) are called Normal Equations.

Substitute the Value of a and b in straight line we get Curve of best fit $y = \hat{a} + \hat{b}x$.

Ques: Fit a curve of $y = a_0x + a_1$ by the PLS.

Note: Regression Equation of y on x : Normal equations are

$$\sum y = na + b \sum x, \quad \sum xy = a \sum x + b \sum x^2.$$

Regression Equation of x on y : Normal Equations.

$$\sum x = n a + b \sum y, \quad \sum xy = a \sum y + b \sum y^2.$$

→ Calculate the regression equations of y on x from the data

Given below taking deviations from actual means of x and y

Price (Rs)	10	12	13	12	16	15
Amount Demanded	40	38	43	45	37	43

Estimate the likely demand when the price is Rs 20.

Sol: Let Price = x , Demand = y .

We have to find demand (y) when price $x=20$.

So we find Regression Line of y on x ,

$$y - \bar{y} = b_{yx} (x - \bar{x})$$

$$\text{Where } \bar{x} = \frac{\sum x}{n}, \bar{y} = \frac{\sum y}{n}, b_{yx} = \frac{\sum xy}{\sum x^2}.$$

x	$x = x - \bar{x}$	x^2	y	$y = y - \bar{y}$	y^2	xy
10	-3	9	40	-1	1	3
12	-1	1	38	-3	9	3
13	0	0	43	2	4	0
12	-1	1	45	4	16	-4
16	3	9	37	-4	16	-12
15	2	4	43	2	4	4
$\sum x = 78$		$\sum x^2 = 24$	$\sum y = 246$			$\sum xy = -6$

$$\bar{x} = \frac{78}{6} = 13, \bar{y} = \frac{246}{6} = 41, b_{yx} = \frac{-6}{24} = -0.25.$$

$$y - 41 = -0.25(x - 13) \Rightarrow y = -0.25x + 44.25.$$

$$\text{When } x = 20, y = -0.25(20) + 44.25 = 39.25.$$

When the price is Rs 20, the likely demand $y = 39.25$.

UNIT-II:PROBABILITY

Lecture Notes

Probability.

Random Experiment: If an experiment is conducted any number of times under essentially identical conditions there is a set of all possible outcomes associated with it. If the result is not certain and is anyone of the several possible outcomes, the experiment is called a Random Trial or Random Experiment. The outcomes are known as elementary events and a set of outcomes is an event.

Equally Likely Events: Events are said to be equally likely when there is no reason to expect anyone of them rather than anyone of the others.

Ex: When a card is drawn from a pack, any card may be obtained. In this trial all the 52 elementary events are ^{equally} likely.

Exhaustive Events: All possible events in any trial are known as Exhaustive events.

Ex: 1. In tossing a coin there are two exhaustive events i.e Head and Tail
2. In throwing a die, there are 6 exhaustive events i.e Getting 1 or 2 or 3 or 4 or 5 or 6.

Mutually Exclusive Events (OR) Disjoint Events: Events are said to be mutually exclusive and equally if the happening of any one of the events in a trial excludes the happening of any one of the others, i.e. If no two or more of the events can happen simultaneously in the same trial.

$$\Rightarrow E_1 \cap E_2 = \emptyset.$$

Probability: In a random experiment, let there be n mutually exclusive and equally likely elementary events. Let E be an event of the experiment. If m elementary events form event E , then the probability of E is defined as

$$P(E) = \frac{m}{n} = \frac{\text{Number of elementary events in } E}{\text{Total number of elementary events in the random experiment}}$$

Let \bar{E} denotes the event of non-occurrence of E . Then the number of elementary events in \bar{E} is $n-m$ and hence the probability of \bar{E} is

$$P(\bar{E}) = \frac{n-m}{n} = \frac{n}{n} - \frac{m}{n} = 1 - \frac{m}{n} = 1 - P(E).$$

$$\therefore P(E) + P(\bar{E}) = 1.$$

Hence $0 \leq P(E) \leq 1$ and $0 \leq P(\bar{E}) \leq 1$.

\bar{E} is also called Complementary event of E .

Note If $P(E) = 1$ the event E is said to be certain event.

If $P(E) = 0$ the event E is called an Impossible event.

Simple Event: An event in a trial that cannot be further split is called a simple event or an elementary event.

Sample Space: The set of all possible simple events in a trial is called a Sample Space for the trial. Each element of a Sample Space is called a Sample point. Any subset of a Sample Space is an event. It is generally denoted by E .

Sample Space is denoted by S .

Probability - Axiomatic Approach:

Def: Let S be a finite sample space. A real valued function P from the power set of S into \mathbb{R} is called a probability function on S if the following three axioms are satisfied.

1. Axiom of positivity: $P(E) \geq 0$ for every subset E of S .

2. Axiom of Certainty: $P(S) = 1$.

3. Axiom of Union: If E_1, E_2 are disjoint subsets of S ,

$$\text{then } P(E_1 \cup E_2) = P(E_1) + P(E_2).$$

Ex: 1. If a coin is tossed out comes

$$S = \{H, T\}.$$

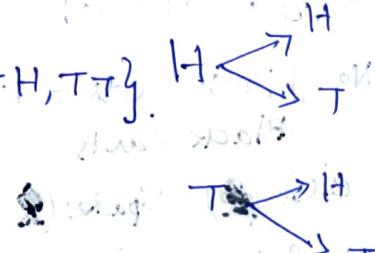
$$\text{No. of outcomes} = 2^1 = 2.$$

2. Two coins are tossed:

$$\text{outcomes } S = \{HH, HT, TH, TT\}.$$

$$\text{No. of outcomes} = 2^2 = 4$$

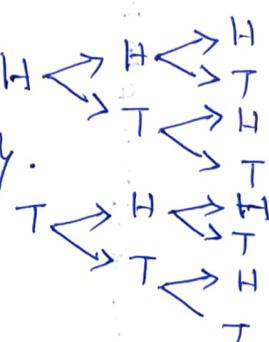
3. 3 coins are tossed.



$$\text{outcomes } S = \{HHH, HHT,$$

$$HTH, HTT, THH, THT, TTH, TTT\}.$$

$$\text{No. of outcomes} = 2^3 = 8$$

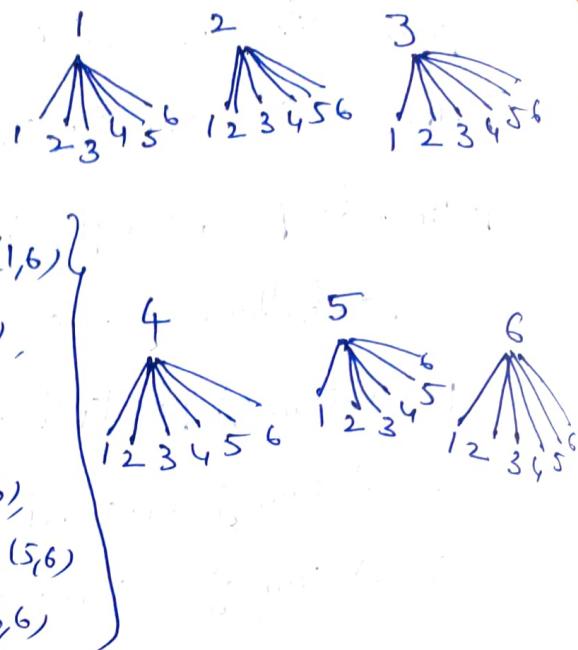


4. If a die is rolled.

$$\text{outcomes } S = \{1, 2, 3, 4, 5, 6\}$$

$$\text{No. of outcomes} = 6^1 = 6$$

5. If Two dies are rolled



$$\text{No. of outcomes} = 6^2 = 36.$$

6. Deck of playing Cards: Total No. of cards = 52.

No. of black cards = 26 [clubs and spades are black]

No. of Red Cards = 26 [Hearts and Diamonds are Red]

$$\text{No. of Face Cards} = 12 \quad \left\{ \begin{array}{l} \text{Jacks, Queens, Kings} \\ \text{Black Cards} \end{array} \right.$$

clubs (♣)	Spades (♠)	Hearts (♥)	Diamonds (♦)
Ace	Ace	Ace	Ace
2	2	2	2
3	3	3	3
4	4	4	4
5	5	5	5
6	6	6	6
7	7	7	7
8	8	8	8
9	9	9	9
10	10	10	10
Jack	Jack	Jack	Jack
Queen	Queen	Queen	Queen
King	King	King	King
<hr/>		<hr/>	
13	13	13	13
<hr/>		<hr/>	
4 x 3 = 12		Face Cards	

Ex: 1. What is the probability for a leap year to have 52 Mondays and 53 Sundays?

Sol: A leap year has 366 days ie 52 weeks and 2 days.

These two days can be any one of the following 7 ways.

- (i) Mon & Tue (ii) Tue & Wed (iii) Wed & Thurs (iv) Thurs & Fri
- (v) Fri & Sat (vi) Sat & Sun (vii) Sun & Mon.

Let E be the event of having 52 Mondays and 53 Sundays in the year.

Total number of possible cases is $n = 7$.

Number of favourable cases to E is $m = 1$. (Sat & Sun is the only favourable case)

$$\therefore P(E) = \frac{m}{n} = \frac{1}{7}$$

→ 2. Determine the probability for each of the following events:

A non-defective bolt will be found if out of 600 bolts already examined 12 were defective.

Sol: The probability of defective bolt $P(D) = \frac{12}{600} = \frac{1}{50} = \frac{m}{n}$.

Where $m = \text{No. of defective bolts} = 12$,

$n = \text{No. of bolts} = 600$.

The prob. of finding a non-defective bolt is

$$P(\bar{D}) = 1 - P(D) = 1 - \frac{1}{50} = \frac{49}{50} = 0.98$$

→ A box contains n tickets marked 1 through n. Two tickets are drawn in succession without replacement. Determine the prob that the number on the tickets are consecutive integers.

Sol: The box contains n tickets marked 1 through n.

The no. of ways in which two tickets can be drawn without replacement = $n(n-1)$.

The total no. of ways of favorable, in which the no. on the tickets are consecutive integers = $n-1$.

$$\text{Required prob} = \frac{n-1}{n(n-1)} = \frac{1}{n}.$$

→ State and prove Addition theorem for three events?

Statement: If A, B, C are three random events then

$$\text{Show that } P(A \cup B \cup C) = P(A) + P(B) + P(C)$$

$$= P(A \cap B) + P(B \cap C) + P(C \cap A) + P(A \cap B \cap C).$$

Proof: We know that addition theorem for two events

$$\therefore P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$$

$$\text{Let } P(A \cup B \cup C) = P(A \cup D) \quad \text{Where } D = B \cup C.$$

$$= P(A) + P(D) - P(A \cap D).$$

$$= P(A) + P(B \cup C) - P[A \cap (B \cup C)] + P(A \cap B \cap C)$$

$$= P(A) + P(B) + P(C) - P(B \cap C) - P[(A \cap B) \cup (A \cap C)]$$

$$= P(A) + P(B) + P(C) - P(B \cap C) - \{P(A \cap B) + P(A \cap C) \\ - P(A \cap B \cap C)\}$$

$$= P(A) + P(B) + P(C) - P(B \cap C) - P(A \cap B) - P(A \cap C)$$

$$+ P(A \cap B \cap C)$$

= R.H.S.

$$\therefore P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C)$$

$$- P(C \cap A) + P(A \cap B \cap C).$$

→ Conditional event: If E_1, E_2 are events of a sample space S and if E_2 occurs after the occurrence of E_1 , then the event of occurrence of E_2 after the event E_1 is called Conditional event of E_2 given E_1 . It is denoted by $\frac{E_2}{E_1}$. Similarly we define $\frac{E_1}{E_2}$.

- Ex: ① Two coins are tossed. The event of getting two tails given that there is at least one tail is a conditional event.
- ② Two unbiased dice are thrown. If the sum of the numbers thrown on them is 7, the event of getting 1 on anyone of them is a conditional event.

→ Conditional probability: If E_1 and E_2 are two events in a sample space S and $P(E_1) \neq 0$, then the probability of E_2 after the event E_1 has occurred is called the Conditional probability of the event of E_2 given E_1 and is denoted by $P(E_2/E_1)$ or $P(\frac{E_2}{E_1})$ and we define $P(E_2/E_1) = \frac{P(E_1 \cap E_2)}{P(E_1)}$.

Similarly we define $P(\frac{E_1}{E_2}) = \frac{P(E_1 \cap E_2)}{P(E_2)}$.

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{\text{No. of elements in } E_1 \cap E_2}{\text{No. of elements in } E_2}$$

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

→ State and prove Multiplication theorem in probability?

Sol: Statement: In a random experiment if E_1, E_2 are two events such that $P(E_1) \neq 0$ and $P(E_2) \neq 0$ then
 $P(E_1 \cap E_2) = P(E_1) \cdot P(E_2/E_1)$ (or)
 $P(E_1 \cap E_2) = P(E_2) \cdot P(E_1/E_2)$.

Proof: Let S be the sample space associated with the random experiment. Let E_1, E_2 be two events of S such that $P(E_1) \neq 0, P(E_2) \neq 0$. Since $P(E_1) \neq 0$, by the definition of conditional probability of E_2 given E_1 ,

$$P(E_2/E_1) = \frac{P(E_1 \cap E_2)}{P(E_1)}.$$

$$\Rightarrow P(E_1 \cap E_2) = P(E_1) \cdot P(E_2/E_1).$$

Since $P(E_2) \neq 0$ $P(E_1/E_2) = \frac{P(E_1 \cap E_2)}{P(E_2)}$

$$\Rightarrow P(E_1 \cap E_2) = P(E_2) \cdot P(E_1/E_2).$$

→ Compound Event: When two or more events occur in conjunction with each other, their joint occurrence is called compound event.

Ex: ① If 2 balls are drawn from a bag containing 4 green, 6 black, 7 white balls, the event of drawing 2 green balls or 2 white balls is a compound event.

→ Independent: If the occurrence of the event E_2 is not affected by the occurrence or non-occurrence of the event E_1 , then the event E_2 is said to be Independent of E_1 , and $P(E_2/E_1) = P(E_2)$.

If $P(E_1) \neq 0$, $P(E_2) \neq 0$ and E_2 is independent of E_1 , then E_1 is independent of E_2 . In this case we say that E_1, E_2 are Mutually Independent or Simply Independent.

→ Dependent event: If the occurrence of the event E_2 is effected by the occurrence of E_1 , then the events E_1, E_2 are dependent and $P(E_2/E_1) \neq P(E_2)$.

→ Theorem: If E_1 and E_2 are independent events of a Sample Space S , then (i) \bar{E}_1 and \bar{E}_2 are independent (ii) $\bar{E}_1 \cup \bar{E}_2$ are independent (iii) $E_1 \rightarrow \bar{E}_2$ are independent.

If A and B are independent events show that (A^c, B^c) , (A, B^c) and (A^c, B) are also independent.

Proof: If E_1, E_2 are independent events then $P(E_1 \cap E_2) = P(E_1)P(E_2)$.

$$\begin{aligned}
 \text{(i)} \quad P(\bar{E}_1 \cap \bar{E}_2) &= P(\bar{E}_1 \cup \bar{E}_2) \\
 &= P(S - (E_1 \cup E_2)). \\
 &= P(S) - P(E_1 \cup E_2) \\
 &= 1 - \{P(E_1) + P(E_2) - P(E_1 \cap E_2)\} \\
 &= 1 - P(E_1) - P(E_2) + P(E_1)P(E_2) \\
 &= (1 - P(E_1)) - P(E_2)[1 - P(E_1)] \\
 &= [1 - P(E_1)][1 - P(E_2)] \\
 &= P(\bar{E}_1)P(\bar{E}_2).
 \end{aligned}$$

∴ \bar{E}_1 and \bar{E}_2 are independent.

$$(ii) P(\bar{E}_1 \cap E_2) = P(E_2 - (E_1 \cap E_2))$$

$$= P(E_2) - P(E_1 \cap E_2)$$

$$= P(E_2) - P(E_1)P(E_2)$$

$$= P(E_2)[1 - P(E_1)]$$

$$= P(E_2)P(\bar{E}_1).$$

$\therefore \bar{E}_1$ and E_2 are independent.

$$(iii) P(E_1 \cap \bar{E}_2) = P(E_1 - (E_1 \cap \bar{E}_2))$$

$$= P(E_1) - P(E_1 \cap \bar{E}_2)$$

$$= P(E_1) - P(E_1)P(\bar{E}_2)$$

$$= P(E_1)[1 - P(\bar{E}_2)]$$

$$= P(E_1)P(\bar{E}_2).$$

$\therefore E_1$ and \bar{E}_2 are independent.

Theorem: If E_1, E_2, E_3 are mutually independent events of a Sample Space S then S . T $E_1 \cup E_2$ and E_3 are also independent events.

Proof: If E_1, E_2, E_3 are independent $\Rightarrow P(E_1 \cap E_2 \cap E_3) = P(E_1)P(E_2)P(E_3)$.

$$P(E_1 \cup E_2 \cap E_3)$$

$$= P(E_1 \cap E_3) \cup (E_2 \cap E_3)$$

$$= P(E_1 \cap E_3) + P(E_2 \cap E_3) - P(E_1 \cap E_2 \cap E_3) \quad \text{Addition Rule}$$

$$= P(E_1)P(E_3) + P(E_2)P(E_3) - P(E_1)P(E_2)P(E_3).$$

$$= P(E_1)P(E_3) + P(E_2)P(E_3) - P(E_1)P(E_2)P(E_3)$$

$$= [P(E_1) + P(E_2) - P(E_1 \cap E_2)]P(E_3)$$

$$= P(E_1 \cup E_2)P(E_3).$$

$\therefore E_1 \cup E_2$ and E_3 are independent events.

→ A card is drawn from a well shuffled pack of cards.

What is the prob that it is either a Spade or an ace?

Sol: Let S be a sample space of all the simple events, $n=52$.

Let A denote the event of getting a Spade and B denote the event of getting an ace.

Then $A \cup B =$ [The event of getting a Spade or an ace] = ?

$A \cap B =$ [The event of getting a Spade and an ace] = 1.

No. of Spades = 13, No. of Aces = 4.

$$\therefore P(A) = \frac{m_1}{n} = \frac{13}{52}, \quad P(B) = \frac{m_2}{n} = \frac{4}{52}, \quad P(A \cap B) = \frac{m_3}{n} = \frac{1}{52}$$

By Addition theorem $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

$$P(A \cup B) = \frac{13}{52} + \frac{4}{52} - \frac{1}{52} = \frac{17}{52} = \frac{17}{52} = \frac{4}{13}$$

→ Three students A, B, C are in running race. A and B have the same prob of winning and each is twice as likely to win as C. Find the prob that B or C wins.

Sol: $A \cup B \cup C = S$ = Sample Space of race.

By data $P(A) = P(B)$ and $P(A) = 2P(C) \rightarrow P(B) = 2P(C) \rightarrow ①$

We have $P(A) + P(B) + P(C) = P(S)$

Sum of the prob = 1.

$$P(A) + P(B) + P(C) = 1 \quad \because A, B, C \text{ are disjoint.} \quad ②$$

$$2P(C) + 2P(C) + P(C) = 1 \quad \because ① \quad ③$$

$$5P(C) = 1 \Rightarrow P(C) = \frac{1}{5} \quad P(A) = \frac{2}{5} \quad P(B) = \frac{2}{5} \quad P(B \cap C) = 0.$$

$$\begin{aligned} \text{The prob that B or C wins} &= P(B \cup C) = P(B) + P(C) - P(B \cap C) \\ &= \frac{2}{5} + \frac{1}{5} - 0 = \frac{3}{5}. \end{aligned}$$

→ From a city 3 news papers A, B, C are being published. A is read by 20%, B is read by 16%. C is read by 14%. both A and B are read by 8%. both A and C are read by 5%. both B and C are read by 4%. and all three A, B, C are read by 2%. What is the percentage of the population that read at least one paper.

Sol: Given that $P(A) = 20\% = \frac{20}{100}$, $P(B) = \frac{16}{100}$, $P(C) = \frac{14}{100}$.

$$P(A \cap B) = \frac{8}{100}, P(A \cap C) = \frac{5}{100}, P(B \cap C) = \frac{4}{100}, P(A \cap B \cap C) = \frac{2}{100}$$

By Addition th by three events

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$

$$= \frac{20}{100} + \frac{16}{100} + \frac{14}{100} - \frac{8}{100} - \frac{5}{100} - \frac{4}{100} + \frac{2}{100}$$

$$= \frac{35}{100}$$

∴ Percentage of the population that read at least one paper

$$\text{Paper} = \frac{35}{100} \times 100 = 35\%$$

→ A bag contains 12 balls numbered from 1 to 12.

If a ball is taken at random. What is the probability of having a ball with a number which is a multiple of either 2 or 3.

Probability of getting a multiple of 2 = $\frac{6}{12}$
 Probability of getting a multiple of 3 = $\frac{4}{12}$

Sol: Number of ways of drawing a ball from the given bag = 12

Number of ways in which it will be a multiple of 2 = 6 {ie } $\{2, 4, 6, 8, 10, 12\}$
 $P(E_1) = \frac{6}{12}$

Number of ways in which the number on the ball is a multiple of 3 = 4
 $P(E_2) = \frac{4}{12}$ {ie } $\{3, 6, 9, 12\}$

Number of ways in which the no. on the ball is multiple of 2 and 3
 $\therefore E_1 \cap E_2 = 2$ {ie } $\{6, 12\} \Rightarrow P(E_1 \cap E_2) = \frac{2}{12}$

By Addition theorem of prob. $P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$
 $P(E_1 \cup E_2) = \frac{6}{12} + \frac{4}{12} - \frac{2}{12} = \frac{10-2}{12} = \frac{8}{12} = \frac{2}{3}$

→ One card is drawn from a regular deck of 52 cards.
What is the prob of card being either red or a king?

Sol: No. of ways drawing one card from a pack of 52 = 52
Number of ways in which it is red = 26.

The prob of drawing red card (E_1) = $P(E_1) = \frac{26}{52}$

No. of ways in which it is a king (E_2) = 4, $P(E_2) = \frac{4}{52}$

No. of ways in which it is a King and red is ($E_1 \cap E_2$) = 2.

$P(E_1 \cap E_2) = \frac{2}{52}$

By addition theorem, $P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$

$\therefore P(E_1 \cup E_2) = \frac{26}{52} + \frac{4}{52} - \frac{2}{52} = \frac{18}{52} = \frac{9}{26}$

Baye's Theorem: Statement

E_1, E_2, \dots, E_n are n mutually exclusive and exhaustive events such that $P(E_i) > 0, i=1, 2, \dots, n$ in a sample space S and A is any other event in S intersecting with every E_i such that $P(A) > 0$.

If E_k is any of the events E_1, E_2, \dots, E_n where $P(E_1), P(E_2), \dots, P(E_n)$ and $P(A|E_1), P(A|E_2), \dots, P(A|E_n)$ are known then

$$P(E_k|A) = \frac{P(E_k) P(A|E_k)}{P(E_1) P(A|E_1) + P(E_2) P(A|E_2) + \dots + P(E_n) P(A|E_n)}$$

Proof: E_1, E_2, \dots, E_n are n events of S such that $P(E_i) > 0$ and $E_i \cap E_j = \emptyset$, for $i \neq j$ where $i, j = 1, 2, \dots, n$. Also E_1, E_2, \dots, E_n are exclusive events of S and A is any other event of S where $P(A) > 0$.

$$\therefore S = E_1 \cup E_2 \cup E_3 \dots \cup E_n \text{ and}$$

$$A = A \cap S = A \cap (E_1 \cup E_2 \cup E_3 \dots \cup E_n).$$

$$= A \cap E_1 \cup A \cap E_2 \cup A \cap E_3 \cup \dots \cup A \cap E_n.$$

Here $A \cap E_1, A \cap E_2, \dots, A \cap E_n$ are mutually exclusive events.

$$\begin{aligned} \text{Then } P(E_k|A) &= \frac{P(E_k \cap A)}{P(A)} = \frac{P(E_k \cap A)}{P[(A \cap E_1) \cup (A \cap E_2) \cup \dots \cup (A \cap E_n)]} \\ &= \frac{P(E_k \cap A)}{P(A \cap E_1) + P(A \cap E_2) + \dots + P(A \cap E_n)}. \end{aligned}$$

$$P(E_k|A) = \frac{P(E_k) P(A|E_k)}{P(E_1) P(A|E_1) + P(E_2) P(A|E_2) + \dots + P(E_n) P(A|E_n)}.$$

Hence the proof of the theorem

→ In a certain college 25% of boys and 10% of girls are studying mathematics. The girls constitute 60% of the students.

(a) What is the prob that mathematics is being studied?

(b) If a student is selected at random and is found to be studying mathematics, find the prob that the student is a girl (c) a boy.

Sol: Given that $P(\text{Boy}) = P(B) = 40\% = \frac{40}{100} = \frac{2}{5}$.

$$P(\text{Girl}) = P(G) = 60\% = \frac{60}{100} = \frac{3}{5}.$$

Prob that mathematics is studied given that the student is a boy
is $P(M/B) = 25\% = \frac{25}{100} = \frac{1}{4}$.

Prob that mathematics is studied given that the student is a girl
is $P(M/G) = 10\% = \frac{10}{100} = \frac{1}{10}$.

(a) Prob that the student studied mathematics

$$P(M) = P(G)P(M/G) + P(B)P(M/B)$$

$$= \frac{3}{5} \cdot \frac{1}{10} + \frac{2}{5} \cdot \frac{1}{4} = \frac{4}{25}.$$

Total Prob Theorem.

(b) By Bayes' Theorem, prob of mathematics student is a girl

$$P(G/M) = \frac{P(G)P(M/G)}{P(M)} = \frac{\frac{3}{5} \cdot \frac{1}{10}}{\frac{4}{25}} = \frac{3}{8}.$$

(c) Prob of mathematics student is a boy is $P(B/M) = \frac{P(B)P(M/B)}{P(M)}$

$$= \frac{\frac{2}{5} \cdot \frac{1}{4}}{\frac{4}{25}} = \frac{5}{8}.$$

→ The chance that doctor A diagnose a disease X correctly is 60%. The chance that a patient will die by his treatment after correct diagnosis is 40%, and the chance of death by wrong diagnosis is 70%. A patient of doctor A, who had 42 disease X, died. What is the chance that his disease was diagnosed correctly.

Sol: Let E_1 be the event that disease X is diagnosed correctly by doctor A and E_2 be the event that a patient of doctor A who has disease X died.

$$\therefore P(E_1) = \frac{60}{100} = 0.6 \Rightarrow P(\bar{E}_1) = 1 - P(E_1) = 1 - 0.6 = 0.4$$

$$P(E_2|E_1) = \frac{40}{100} = 0.4 \Rightarrow P(\bar{E}_2|\bar{E}_1) = 1 - 0.4 = \frac{70}{100} = 0.7.$$

By Bayes' Theorem

$$P(E_1|E_2) = \frac{P(E_1) P(E_2|E_1)}{P(E_1) P(E_2|E_1) + P(\bar{E}_1) P(\bar{E}_2|\bar{E}_1)}.$$

$$P(E_1|E_2) = \frac{0.6 \times 0.4}{0.6 \times 0.4 + 0.4 \times 0.7} = \frac{6}{13}.$$

→ A bag A contains 2 white and 3 red balls and a bag B contains 4 white and 5 red balls. One ball is drawn at random from one of the bags and it is found to be red. Find the prob that the red ball drawn is from bag B?

Sol: Given that

Bags	White	Red	Total
A	2	3	5
B	4	5	9

There are two bags, $P(A) = \frac{1}{2}$, $P(B) = \frac{1}{2}$.

Let R denote the event of drawing a red ball.

The prob to draw a red ball from bag A is $P(R|A) = \frac{3}{5}$.

Similarly $P(R|B) = \frac{5}{9}$.

Prob of Red ball is $P(R) = P(A)P(R|A) + P(B)P(R|B)$

$$= \frac{1}{2} \cdot \frac{3}{5} + \frac{1}{2} \cdot \frac{5}{9} = \frac{3}{10} + \frac{5}{18} = \frac{54+50}{180} = \frac{104}{180}$$

Prob that ball is red, drawn from bag B

$$P(B|R) = \frac{P(B)P(R|B)}{P(R)} = \frac{\frac{1}{2} \cdot \frac{5}{9}}{\frac{104}{180}} = \frac{5}{18} \times \frac{180}{104} = \frac{25}{52}$$

→ Suppose 5 men out of 100 and 25 women out of 10,000 are colour blind. A colour blind person is chosen at random. What is the prob of the person being a male? (Assume male and female are equal in numbers)

Sol: Prob that the chosen person is male $P(M) = \frac{1}{2}$

Prob that the chosen person is female $P(W) = \frac{1}{2}$

Given that 5 men out of 100 are colour blind i.e $P(B|M) = \frac{5}{100}$

Also 25 women out of 10,000 are colour blind $P(B|W) = \frac{25}{10,000}$

Where B is colour blind person.

$$\therefore \text{Prob of Colour blind person } P(B) = P(M)P(B|M) + P(W)P(B|W)$$

$$= \frac{1}{2} \times \frac{5}{100} + \frac{1}{2} \times \frac{25}{10,000} = \frac{5}{200} + \frac{25}{20,000} = \frac{1}{40} + \frac{1}{800}$$

$$= \frac{20+1}{800} = \frac{21}{800} = 0.0262$$

Prob that the person being a male, he is colour blind

$$P(M|B) = \frac{P(M)P(B|M)}{P(B)} = \frac{\frac{1}{2} \times \frac{5}{100}}{0.0262} = \frac{0.025}{0.0262} = 0.95$$

Random Variables.

Def: A real variable x whose value is determined by the outcome of a random experiment is called a Random Variable.
 A random variable x can also be regarded as a real-value function defined on the sample space S of a random experiment such that for each point ω of the sample space, $f(\omega)$ is the probability of occurrence of the event represented by x .

Types of Random variables: There are two types,

1. Discrete Random Variable. (2) Continuous Random Variable.

→ **Discrete Random Variable:** A random variable X which can take only a finite number of discrete values in an interval of domain is called a discrete Random Variable.

→ **Continuous Random Variable:** A random variable X which can take values Continuously i.e. which takes all possible values in a given interval is called a Continuous random variable.

Probability Function of A Discrete Random Variable:

If for a discrete random variable X , there is a real valued function $p(x)$ such that $p(X=x) = p(x)$ then $p(x)$ is called probability function (OR) probability density function.

Properties of a Probability Function: If $p(x)$ is a probability function of a random variable X , then it possesses the following properties: (i) $p(x) \geq 0$, for all x .

(2) $\sum p(x) = 1$. Sum of the probabilities is equal to one.

(3) $p(x)$ cannot be negative for any value of x .

Probability Density Function: The probability density function

$f_x(x)$ is defined as the derivative of the probability distribution function, $F_x(x)$ of the random variable x .

$$\Rightarrow f_x(x) = \frac{d}{dx} [F_x(x)].$$

Distribution Function: If $p(x)$ i.e. $f(x)$ is the probability function or probability distribution then the value of $\sum_{x=0}^{\infty} p(x)$ i.e. $\sum_{x=0}^{\infty} f(x)$

denoted by $F(x)$ is called the Cumulative Distribution Function or Distribution Function.

Mean or Expected Value of x is $E(x)$ or μ .

$$\therefore \mu = E(x) = \sum_{x=0}^n x p(x).$$

$$\text{Variance } \sigma^2 = E(x^2) - [E(x)]^2$$

$$= \sum_{x=0}^n x^2 p(x) - \left[\sum_{x=0}^n x p(x) \right]^2$$

$$\text{Standard Deviation } \sigma = \sqrt{\sum(x^2) - [E(x)]^2}.$$

Theorem: If x is a random variable, then $V(ax+b) = a^2 V(x)$ where $V(x)$ is variance of x and a, b are constants.

Proof: Let $y = ax + b \rightarrow \textcircled{1}$

$$\text{Then } E(y) = E(ax+b) = a E(x) + b \rightarrow \textcircled{2}$$

$$\textcircled{1} - \textcircled{2} \Rightarrow y - E(y) = a[x - E(x)]$$

Squaring on both sides and take expectation on both sides, we get

$$E[(y - E(y))^2] = a^2 E[(x - E(x))^2]$$

$$\text{i.e. } V(y) = a^2 V(x).$$

$$\Rightarrow V(ax+b) = a^2 V(x).$$

Note: (i) If $b = 0$ then $V(ax) = a^2 V(x)$

(ii) If $a = 0$ then $V(b) = 0$

(iii) If $a = 1$ then $V(x+b) = V(x)$

Theorem: If x and y are any two random variables, then

$$E(x+y) = E(x) + E(y), \text{ provided } E(x) \text{ and } E(y) \text{ exist.}$$

Sol: Let x assume the values x_1, x_2, \dots, x_n and y assume the values y_1, y_2, \dots, y_m , then by defn.

$$E(x) = \sum_{i=1}^n p_i x_i \text{ and } E(y) = \sum_{j=1}^m p_j y_j.$$

$$\text{let } P_{ij} = P(x=x_i \cap y=y_j) = p(x_i, y_j).$$

The sum $(x+y)$ is also a random variable which can take $m \times n$ values $(x_i + y_j)$, $i=1, 2, \dots, n$, $j=1, 2, \dots, m$.

$$\text{By defn } E(x+y) = \sum_{i=1}^n \sum_{j=1}^m P_{ij} (x_i + y_j).$$

$$= \sum_{i=1}^n \sum_{j=1}^m P_{ij} x_i + \sum_{j=1}^m \sum_{i=1}^n P_{ij} y_j.$$

$$= \sum_{i=1}^n \left[x_i \sum_{j=1}^m P_{ij} \right] + \sum_{j=1}^m \left[y_j \sum_{i=1}^n P_{ij} \right]$$

$$= \sum_{i=1}^n x_i p_i + \sum_{j=1}^m y_j p_j$$

$$\underline{E(x+y) = E(x) + E(y)}$$

Ques: If $X = 3x_1 + 4x_2 + 6x_3 + 8x_4 + 10x_5$ and $P(x_i) = \frac{1}{5}$ for all i .
Find $E(X)$ and $V(X)$.

→ A random variable X has the following probability function

x_i	-3	-2	-1	0	1	2	3
$p(x_i)$	K	0.1	K	0.2	$2K$	0.4	$2K$

Find (i) K (ii) Mean (iii) Variance.

Sol: We know that Sum of the probabilities is equal to 1.

$$\Rightarrow \sum p(x_i) = 1 \Rightarrow K + 0.1 + K + 0.2 + 2K + 0.4 + 2K = 1$$

$$\Rightarrow 6K + 0.7 = 1 \Rightarrow 6K = 1 - 0.7 = 0.3$$

$$\therefore K = \frac{0.3}{6} = \frac{3}{60} = \frac{1}{20}$$

(i) Mean $\mu = \sum x p(x)$

$$\begin{aligned}\mu &= (-3)K + (-2)(0.1) + (-1)(K) + 0(0.2) + 1(2K) + 2(0.4) \\ &= -3K - 0.2 - K + 0 + 2K + 0.8 + 6K \\ &= 4K + 0.6 = 4\left(\frac{1}{20}\right) + 0.6 = \frac{1}{5} + 0.6 = 0.2 + 0.6 \\ \therefore \mu &= 0.8\end{aligned}$$

(iii) Variance $\sigma^2 = [\sum x^2 p(x)] - \mu^2$

$$\begin{aligned}&= [(-3)^2 K + (-2)^2 (0.1) + (-1)^2 K + 0^2 (0.2) + 1^2 (2K) + 2^2 (0.4) + 3^2 (2K)] \\ &= [9K + 0.4 + K + 0 + 2K + 0.4(4) + 18K] - (0.8)^2 \\ &= [30K + 2] - 0.64 \\ &= [30\left(\frac{1}{20}\right) + 2] - 0.64 \\ &= (1.5 + 2) - 0.64 \\ &= 3.5 - 0.64 \\ &= 2.86\end{aligned}$$

Variance = 2.86

Continuous Probability distributions.

Mean $\mu = E(x) = \int_a^b x f(x) dx$.

Median (M) $\leftarrow \int_a^M f(x) dx = \int_M^b f(x) dx = \frac{1}{2}$. Solving for M we get Median.

Mode = Mode is the value of x for which $f(x)$ is maximum.

Mode is thus given by $f'(x) = 0$ and $f''(x) < 0$ for $a < x < b$.

Variance $\sigma^2 = \left[\int_a^b x^2 f(x) dx \right] - \mu^2$.

Mean deviation about μ is $\int_{-\infty}^{\infty} |x - \mu| f(x) dx$.

→ If a random variable has the prob density function $f(x)$ as

$$f(x) = \begin{cases} 2e^{-2x}, & \text{for } x > 0 \\ 0, & \text{for } x \leq 0. \end{cases}$$

Find probability,

- (i) between 1 and 3 (ii) greater than 0.5.

Sol: (i) Prob between 1 and 3 is given by $p(1 \leq x \leq 3) = \int_1^3 f(x) dx$

$$= \int_1^3 2e^{-2x} dx = 2 \left[\frac{e^{-2x}}{-2} \right]_1^3 = -1 \left[e^{-6} - e^{-2} \right] = e^2 - e^6.$$

(ii) Prob greater than 0.5 is $p(x \geq 0.5) = \int_{0.5}^{\infty} f(x) dx$

$$= \int_{0.5}^{\infty} 2e^{-2x} dx = 2 \left[\frac{e^{-2x}}{-2} \right]_{0.5}^{\infty} = -1 \left[e^{-4} - e^{-2(0.5)} \right] = -1 \left[0 - e^{-2} \right]$$

$$= \frac{1}{e^2}.$$

→ If the probability density of a random variable is given by

$$f(x) = \begin{cases} k(1-x^2), & 0 < x < 1 \\ 0, & \text{otherwise} \end{cases}$$

Find the value of k and the prob

that a random variable having this prob density units take on a value

- (i) between 0.1 and 0.2 (ii) greater than 0.5.

Sol: Given that $f(x) = \begin{cases} k(1-x^2), & 0 < x < 1 \\ 0, & \text{otherwise} \end{cases}$

We know that $\int_{-\infty}^{\infty} f(x) dx = 1$. [\because Sum of the prob = 1.]

$$\begin{aligned}
 & \text{ie} \int_{-\infty}^0 f(x)dx + \int_0^1 f(x)dx + \int_1^\infty f(x)dx = 1 \\
 & \Rightarrow 0 + \int_0^1 k(1-x^2)dx + 0 = 1 \Rightarrow k \int_0^1 (1-x^2)dx = 1 \\
 & \Rightarrow k \left\{ \left[x \right]_0^1 - \left[\frac{x^3}{3} \right]_0^1 \right\} = 1 \Rightarrow k \left[(1-0) - \left(\frac{1}{3}-0 \right) \right] = 1 \\
 & \Rightarrow k(1-\frac{1}{3}) = 1 \Rightarrow k(\frac{2}{3}) = 1 \Rightarrow k = \frac{3}{2} \quad k = \frac{3}{2}
 \end{aligned}$$

$$\begin{aligned}
 & \text{(i) Prob between } 0.1 \text{ and } 0.2 \text{ is } p(0.1 < x < 0.2) = \int_{0.1}^{0.2} f(x)dx \\
 & = \int_{0.1}^{0.2} k(1-x^2)dx = \frac{3}{2} \int_{0.1}^{0.2} (1-x^2)dx = \frac{3}{2} \left\{ x - \frac{x^3}{3} \right\}_{0.1}^{0.2} \\
 & = \frac{3}{2} \left[(0.2) - (0.1) \right] - \frac{1}{3} \left[(0.2)^3 - (0.1)^3 \right] \\
 & = \frac{3}{2} \left[0.1 - \frac{1}{3}(0.008 - 0.001) \right] = \frac{3}{2} \left[0.1 - \frac{0.007}{3} \right] = 0.2965
 \end{aligned}$$

$$\begin{aligned}
 & \text{(ii) Prob that greater than } 0.5 \text{ is } p(x > 0.5) = \int_{0.5}^\infty f(x)dx \\
 & = \int_{0.5}^1 f(x)dx + \int_1^\infty f(x)dx = \int_{0.5}^1 k(1-x^2)dx + 0 \quad k = \frac{3}{2} \\
 & = \frac{3}{2} \left\{ x - \frac{x^3}{3} \right\}_{0.5}^1 = \frac{3}{2} \left\{ (1-0.5) - \left(\frac{1-(0.5)^3}{3} \right) \right\} \\
 & = \frac{3}{2} \left\{ 0.5 - \left(\frac{1-0.125}{3} \right) \right\} = \frac{3}{2} \left\{ 0.5 - \frac{0.875}{3} \right\} \\
 & = \frac{3}{2} \left[0.5 - 0.2916 \right] = 1.5(0.2084) = 0.3126
 \end{aligned}$$

UNIT-III:PROBABILITY DISTRIBUTION

Lecture Notes

Unit-III. Probability Distributions.

Bernoulli's Distribution: A random variable x which takes two values 0 and 1 with probability q and p respectively ie $P(x=0)=q$ and $P(x=1)=p$, $q=1-p$ is called a Bernoulli's discrete random variable and is said to have a Bernoulli's distribution.

The probability function of Bernoulli's distribution can be

written as $p(x) = p^x q^{1-x} = p^x (1-p)^{1-x}$, $x=0,1$.

Def: Suppose, associated with a random trial there is an event called "Success" and the complementary event called "Failure". Let the probability for Success be p and probability for failure be q . Suppose the random trials are performed n times under identical conditions. These are called Bernoullian trials.

Binomial Distribution.

Binomial distribution was discovered by James Bernoulli in the year 1700 and it is a discrete probability distribution. Let us visualise a conceptual or practical situation where a trial or an experiment results in only two outcomes, say 'Success' and 'Failure', the result of one trial does not influence the result of next trial, and the probability of success at each trial is the same from trial to trial.

Some of such situations are

1. Tossing a coin - Head, or Tail
2. Birth of a baby - Girl or Boy
3. Auditing a bill - Contains an error or not
4. An advertisement on TV - Recalled by viewer or not

Conditions for the applicability of a Binomial distribution are

(or) **BERNOULLI** conditions

1. There are n independent trials
2. Each trial has only two possible outcomes
3. The probabilities of two outcomes remain constant.

Def: A random variable X has a Binomial distribution if it assumes only non-negative values and its probability density function is given by

$$P(X=r) = p(r) = \begin{cases} nCr p^r q^{n-r}, & r=0, 1, 2, \dots, n, q=1-p, \\ 0, & \text{otherwise} \end{cases}$$

Binomial distribution function is given by

$$F(x) = P(X \leq x) = \sum_{r=0}^x nCr p^r q^{n-r}$$

Examples of Binomial Distribution:

1. The number of defective bolts in a box containing n bolts
2. The number of machines lying idle in a factory having n machines.
3. The number of post-graduates in a group of n men
4. The number of oil wells yielding natural gas in a group of n wells test drilled.

Mean of the Binomial distribution :-

The Binomial probability distribution is given by

$$P(r) = nC_r p^r q^{n-r}, r=0, 1, 2, \dots, n \text{ and } q = 1-p.$$

Mean of x is $\mu = E(x) = \sum_{r=0}^n r P(r).$

$$= 0 \times q^n + 1 \times nC_1 p q^{n-1} + 2 \times nC_2 p^2 q^{n-2} + \dots + n p^n.$$

$$= npq^{n-1} + 2 \frac{n(n-1)}{2!} p^2 q^{n-2} + 3 \cdot \frac{n(n-1)(n-2)}{3!} p^3 q^{n-3} + \dots + np^n.$$

$$= np \left[q^{n-1} + (n-1)pq^{n-2} + \frac{(n-1)(n-2)}{2!} p^2 q^{n-3} + \dots + p^{n-1} \right]$$

$$= np(q+p)^{n-1}$$

By Using Binomial theorem.

$$\therefore \mu = np(1)^{n-1} = np \quad \because p+q=1.$$

Hence the Arithmetic mean of the Binomial distribution = np .

Variance of the Binomial distribution :

$$\text{Variance } \sigma^2 = E(x^2) - [E(x)]^2$$

$$= \sum_{r=0}^n r^2 P(r) - \mu^2 = \sum_{r=0}^n [r(r-1) + r] P(r) - \mu^2 \quad \text{Add and Subtract } r$$

$$= \sum_{r=0}^n r(r-1) nC_r p^r q^{n-r} + \sum_{r=0}^n r P(r) - \mu^2 : \sum_{r=0}^n r P(r) = \mu$$

$$= [2 \cdot nC_2 p^2 q^{n-2} + 3 \cdot 2 \cdot nC_3 p^3 q^{n-3} + \dots + n(n-1) p^n] + \mu - \mu^2$$

$$= \left[\frac{n(n-1)}{2!} p^2 q^{n-2} + 6 \cdot \frac{n(n-1)(n-2)}{3!} p^3 q^{n-3} + \dots + n(n-1) p^n \right] + \mu - \mu^2$$

$$= n(n-1) p^2 \left[q^{n-2} + (n-2)pq^{n-3} + \frac{(n-2)(n-3)}{2!} p^2 q^{n-4} + \dots + p^{n-2} \right] + \mu - \mu^2$$

$$= n(n-1) p^2 (q+p)^{n-2} + \mu - \mu^2 \quad \text{by Binomial Theorem.}$$

$$= n(n-1) p^2 (1)^{n-2} + \mu - \mu^2 \quad \therefore q+p=1.$$

$$= n(n-1) p^2 + np - (\mu)^2 \quad \therefore \mu = np = \text{mean.}$$

$$= n^2 p^2 - np^2 + np = np(n-p)$$

$$= -np^2 + np = np(1-p) = npq$$

∴ Variance of the Binomial distribution, $\sigma^2 = npq$

Hence the Standard Deviation of the Binomial Distribution, $\sigma = \sqrt{npq}$

Mode of the Binomial Distribution:

Mode of the Binomial distribution is the value of r at which $p(r)$ is maximum.

$$\text{Mode} = \begin{cases} \text{integral part of } (n+1)p, \text{ if } (n+1)p \text{ is not an integer} \\ (n+1)p \text{ and } (n+1)p-1, \text{ if } (n+1)p \text{ is an integer.} \end{cases}$$

Recurrence Relation for the Binomial distribution:

$$\text{We know that } p(r) = nCr p^r q^{n-r}$$

$$p(r+1) = nCr_{r+1} p^{r+1} q^{n-r-1}$$

$$\textcircled{2} \div \textcircled{1} \text{ gives}$$

$$\frac{p(r+1)}{p(r)} = \frac{nCr_{r+1}}{nCr_r} \cdot \frac{p^{r+1} q^{n-r-1}}{p^r q^{n-r}} = \frac{n-r}{r+1} \cdot \frac{p}{q}$$

$$\therefore p(r+1) = \frac{(n-r)p}{(r+1)q} p(r).$$

Binomial Frequency distribution: If n independent trials constitute one experiment and this experiment is repeated N times, then the frequency of r successes is $N \cdot nCr p^r q^{n-r}$. Since the probabilities of $0, 1, 2, \dots, n$ success in n trials are given by the terms of the binomial expansion of $(p+q)^n$, therefore in N sets of n trials the theoretical frequencies of $0, 1, 2, \dots, n$ successes will be given by the terms of expansion of $N(p+q)^n$. The possible number of successes and their frequencies is called a Binomial Freq. distribution.

→ A fair coin is tossed six times. Find the probability of getting four heads.

Sol: We have $p = \text{probability of getting a head} = \frac{1}{2}$.

$q = \text{prob. of getting a tail} = \frac{1}{2}$.

$n = \text{no. of tosses} = 6$, $r = \text{no. of heads} = 4$.

We know that by Binomial distribution, $P(r) = nCr p^r q^{n-r}$.

$$\therefore P(4) = 6C_4 \left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right)^{6-4} = \frac{6!}{2!4!} \left(\frac{1}{2}\right)^6 = \frac{6 \times 5 \times 4!}{2 \times 4!} \left(\frac{1}{2}\right)^6 \\ = \frac{6 \times 5}{2} \left(\frac{1}{2}\right)^6 = \frac{15}{64} = 0.2344.$$

Probability of getting 4 heads if a coin is tossed six times = 0.2344.

→ Ten coins are thrown simultaneously. Find the probability of getting at least (i) Seven heads (ii) Six heads.

Sol: Let $p = \text{probability of getting a head} = \frac{1}{2}$

$q = \text{probability of not getting a head} = \frac{1}{2}$.

$n = \text{no. of tosses} = 10$, $r = \text{no. of heads} = ?$

The probability of getting r heads in a throw of 10 coins are

$$P(r) = 10Cr \left(\frac{1}{2}\right)^r \left(\frac{1}{2}\right)^{10-r}, \quad r=0, 1, 2, \dots, n. \quad \log_6 = \frac{10 \times 9 \times 8 \times 7 \times 6 \times 5}{8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1} = 210$$

(i) Probability of getting at least seven heads is given by

$$P(X \geq 7) = P(X=7) + P(X=8) + P(X=9) + P(X=10).$$

$$= 10C_7 \left(\frac{1}{2}\right)^7 \left(\frac{1}{2}\right)^{10-7} + 10C_8 \left(\frac{1}{2}\right)^8 \left(\frac{1}{2}\right)^{10-8} + 10C_9 \left(\frac{1}{2}\right)^9 \left(\frac{1}{2}\right)^{10-9} + 10C_{10} \left(\frac{1}{2}\right)^{10}.$$

$$= \left(\frac{1}{2}\right)^{10} [10C_7 + 10C_8 + 10C_9 + 10C_{10}] = \frac{1}{2^{10}} [120 + 45 + 10 + 1] = \frac{176}{1024} = 0.1719$$

ii) Prob. of getting at least 6 heads, $P(X \geq 6) = P(X=6) + P(X=7) + P(X=8) + P(X=9) + P(X=10)$

$$= 10C_6 \left(\frac{1}{2}\right)^6 \left(\frac{1}{2}\right)^{10-6} + [10C_7 \left(\frac{1}{2}\right)^7 \left(\frac{1}{2}\right)^{10-7} + 10C_8 \left(\frac{1}{2}\right)^8 \left(\frac{1}{2}\right)^{10-8} + 10C_9 \left(\frac{1}{2}\right)^9 \left(\frac{1}{2}\right)^{10-9} + 10C_{10} \left(\frac{1}{2}\right)^{10}]$$

$$= \frac{1}{2^{10}} (10C_6) + 0.1719 = \frac{210}{1024} + 0.1719 = 0.2050 + 0.1719 = 0.3767.$$

→ Seven Coins are tossed and the number of heads are noted.
The experiment is repeated 128 times and the following distribution is obtained.

No. of heads	0	1	2	3	4	5	6	7	Total
Frequency	7	6	19	35	30	23	7	1	128

Fit a binomial distribution

- (a) The coin is unbiased.
- (b) The nature of the coin is not known.

Sol: (a) If the coin is unbiased then

$P = \text{prob of head} = \frac{1}{2}$, $q = \frac{1}{2}$ and $n = \text{no. of coins} = 7$.

$$N = \sum f = 7 + 6 + 9 + 35 + 30 + 23 + 7 + 1 = 128$$

By Binomial distribution, $p(r) = n p^r q^{n-r}$

We have the recurrence relation $p(r+1) = \frac{(n-r)p}{(r+1)q} p(r)$

$$p(r+1) = \frac{n-r}{r+1} p(r) \quad n=7, P=\frac{1}{2}, q=\frac{1}{2}, P/q=1.$$

$$P(0) = 7 C_0 (\frac{1}{2})^0 (\frac{1}{2})^{7-0} = (1)(1)(\frac{1}{2})^7 = \frac{1}{2^7}.$$

No. of heads x	Observed frequency f	Probability $p(r)$	Expected or Theoretical frequency $f(r) = N p(r)$
-------------------	----------------------------	-----------------------	--

$$0 \quad 7 \quad P(0) = \frac{1}{2^7} \quad f(0) = 128 p(0) = 128 \times \frac{1}{2^7} = 1.$$

$$1 \quad 6 \quad P(1) = \frac{7}{2^7} \quad f(1) = 128 \times \frac{7}{2^7} = 7$$

$$2 \quad 19 \quad P(2) = \frac{7-1}{2^7} p(1) \\ = \frac{6}{2} \times \frac{7}{2^7} = \frac{21}{2^7} \quad f(2) = 128 \times \frac{21}{2^7} = 21$$

$$3 \quad 35 \quad P(3) = \frac{35}{2^7} \quad f(3) = 128 \times \frac{35}{2^7} = 35$$

$$4 \quad 30 \quad P(4) = \frac{35}{2^7} \quad f(4) = 128 \times \frac{35}{2^7} = 35$$

$$5 \quad 23 \quad P(5) = \frac{21}{2^7} \quad f(5) = 128 \times \frac{21}{2^7} = 21$$

$$6 \quad 7 \quad P(6) = \frac{7}{2^7} \quad f(6) = 128 \times \frac{7}{2^7} = 7$$

$$7 \quad 1 \quad P(7) = \frac{1}{2^7} \quad f(7) = 128 \times \frac{1}{2^7} = 1$$

$$\text{Sum of expected frequencies} = 1 + 7 + 21 + 35 + 35 + 21 + 7 + 1 \\ = 128$$

(b) The nature of the coin is not known.

Given that no. of coins $n = 7$, $N = \sum f = 128$.

$$\text{Mean } \mu = \frac{\sum fx}{\sum f} = \frac{0 \times 7 + 1 \times 6 + 2 \times 19 + 3 \times 35 + 4 \times 30 + 5 \times 23 + 6 \times 7 + 7 \times 1}{128}$$

$$= \frac{6 + 38 + 105 + 120 + 115 + 142 + 7}{128} = \frac{433}{128} = 3.383$$

$$\therefore \text{Mean } np = \frac{3.383}{128} = 7(1)(p) = 3.383 \Rightarrow p = \frac{3.383}{7} = 0.4833$$

$$q = 1 - p = 1 - 0.4833 = 0.5167 \approx 0.5167$$

Hence the Binomial distribution to be fitted is $N(q, p)^n$

$$\text{By B.D. } = 128(0.5167 + 0.4833)^7$$

$$P(r) = nCr P^r q^{n-r}, P(0) = 7C_0 (0.5167)^0 (0.4833)^7$$

$$= (1)(1)(0.4833)^7 = 0.0061 \quad n=7$$

$$P(r+1) = \frac{(7-r)(0.5167)^r}{(r+1)(0.4833)} P(r), \quad P(r) = \frac{(7-r)}{(r+1)} (1.0691) P(r).$$

No. of Heads	Observed frequency	Probability $P(r)$	Expected or Theoretical frequency $f(r) = N P(r)$
0	7	$P(0) = 0.0061$	$f(0) = 128 \times P(0) = 128 \times 0.0061 = 0.78$
1	6	$P(1) = \frac{(7-0)}{(0+1)} (1.0691)(0.0061) = 0.0456$	$f(1) = 128 \times 0.0456 = 5.8$
2	19	$P(2) = \frac{(7-1)}{(1+1)} (1.0691)(0.0061) = 0.0915$	$f(2) = 128 \times 0.0915 = 11.5$
3	35	$P(3) = \frac{(7-2)}{(2+1)} (1.0691)(0.0061) = 0.108$	$f(3) = 128 \times 0.108 = 1.39$
4	30	$P(4) = \frac{(7-3)}{(3+1)} (1.0691)(0.0061) = 0.065$	$f(4) = 128 \times 0.065 = 0.83$
5	23	$P(5) = \frac{(7-4)}{(4+1)} (1.0691)(0.0061) = 0.029$	$f(5) = 128 \times 0.029 = 0.508$
6	7	$P(6) = \frac{(7-5)}{(5+1)} (1.0691)(0.0061) = 0.002$	$f(6) = 128 \times 0.002 = 0.27$
7	1	$P(7) = \frac{(7-6)}{(6+1)} (1.0691)(0.0061) = 0.0009$	$f(7) = 128 \times 0.0009 = 0.11$

Poisson Distribution

Def: A random variable x is said to follow a Poisson distribution if it assumes only non-negative values and its probability density function is given by

$$P(x, \lambda) = P(x=x) = \begin{cases} \frac{e^{-\lambda} \lambda^x}{x!}; & x=0, 1, 2, 3, \dots \\ 0, & \text{otherwise} \end{cases}$$

Here $\lambda > 0$ is called the parameter of the distribution.

Examples of Poisson distribution :-

1. The number of defective electric bulbs manufactured by a reputed company.
2. The no. of telephone calls per minute at a switch board.
3. The no. of cars passing a certain point in one minute.
4. The no. of printing mistakes per page in a large text.
5. The no. of particles emitted by a radio-active substance.
6. The no. of persons born blind per year in a large city.

Conditions of Poisson Distribution:-

The Poisson distribution is used under the following conditions

1. The variable (number of occurrences) is a discrete variable.
2. The occurrences are rare.
3. The number of trials (n) is large.
4. The probability of success (p) is very small (very close to zero).
5. $np = \lambda$ is finite.

Derivation of the poisson distribution:

The poisson distribution can be derived as a limiting case of the Binomial distribution under the following conditions:

- (i) p , the prob of occurrence of the event is very small
- (ii) n is very very large, where n is number of trials, i.e. $n \rightarrow \infty$
- (iii) np is a finite quantity, say $np = \lambda$, then λ is called the parameter of the poisson distribution.

In the Binomial distribution the probability $P(r)$ or $p(x=r)$ of r successes in a series of independent trials is given by

$$P(r) = {}^n C_r p^r q^{n-r} = {}^n C_r p^r (1-p)^{n-r}$$

$$= \frac{n(n-1)(n-2)\dots(n-r+1)}{r!} p^r \cdot \frac{(1-p)^n}{(1-p)^r} \rightarrow ①$$

$$\text{Put } np = \lambda \Rightarrow n = \frac{\lambda}{p}$$

$$\text{Hence } ① \text{ becomes } P(r) = \frac{\frac{\lambda}{p}(\frac{\lambda}{p}-1)(\frac{\lambda}{p}-2)\dots(\frac{\lambda}{p}-r+1)}{r!} \cdot p^r \cdot \frac{(1-p)^n}{(1-p)^r}$$

$$= \frac{\lambda(\lambda-p)(\lambda-2p)\dots[\lambda-(r-1)p]}{r! p^r} \cdot \frac{p^r (1-p)^n}{(1-p)^r}$$

$$= \frac{\lambda(\lambda-p)(\lambda-2p)\dots[\lambda-(r-1)p]}{r!} \cdot \frac{(1-\frac{\lambda}{n})^n}{(1-\frac{\lambda}{n})^r} \rightarrow ②$$

As $n \rightarrow \infty$, $p \rightarrow 0$. So that, $np = \lambda$ we have

$$P(r) = \frac{\lambda \cdot \lambda \dots \lambda}{r!} \text{ (r factors)} \quad \frac{\lambda}{n} = p$$

$$= \frac{\lambda^r}{r!} \underset{n \rightarrow \infty}{\text{LT}} (1 - \frac{\lambda}{n})^n \quad \underset{n \rightarrow \infty}{\text{LT}} (1 - \frac{\lambda}{n})^n \cdot \underset{p \rightarrow 0}{\text{LT}} \frac{1}{(1-p)^r}$$

$$= \frac{\lambda^r}{r!} e^{-\lambda} \quad \left[\because \underset{p \rightarrow 0}{\text{LT}} \frac{1}{(1-p)^r} = \frac{1}{1} = 1 \text{ (given)} \right]$$

$$\therefore P(r) = \frac{e^{-\lambda} \lambda^r}{r!} = \text{Probability of } r \text{ success}$$

This is called the poisson distribution, where, $r = 0, 1, 2, \dots$

Mean of the poisson distribution:

$$\begin{aligned}
 \text{Mean} = E(X) &= \sum_{x=0}^{\infty} x p(x) = \sum_{x=0}^{\infty} x e^{-\lambda} \frac{\lambda^x}{x!} \\
 &= e^{-\lambda} \sum_{x=1}^{\infty} \frac{\lambda^x}{(x-1)!} \quad \because x! = x(x-1)! \\
 &= e^{-\lambda} \sum_{y=0}^{\infty} \frac{\lambda^{y+1}}{y!} \quad \text{putting } y = x-1 \\
 &= \lambda e^{-\lambda} \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} = e^{-\lambda} \lambda e^{\lambda} \quad \therefore \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} = e^{\lambda}
 \end{aligned}$$

$E(X) = \lambda$, ($\lambda = np$) is Mean of poisson distribution.

Variance of poisson distribution:

$$\text{Variance } \sigma^2 = V(X) = E(X^2) - [E(X)]^2$$

$$\begin{aligned}
 &= \sum_{x=0}^{\infty} x^2 p(x) - \lambda^2 \quad \because \lambda = E(X) = \text{Mean of P.D} \\
 &= \sum_{x=0}^{\infty} x^2 \frac{e^{-\lambda} \lambda^x}{x!} - \lambda^2 = e^{-\lambda} \sum_{x=0}^{\infty} \frac{x \lambda^x}{(x-1)!} - \lambda^2 \quad \because x! = x(x-1)! \\
 &= \left[e^{-\lambda} \sum_{x=1}^{\infty} \frac{[(x-1)+1] \lambda^x}{(x-1)!} \right] - \lambda^2 \\
 &= e^{-\lambda} \left[\sum_{x=1}^{\infty} \frac{(x-1) \cdot \lambda^x}{(x-1)(x-2)!} + \sum_{x=1}^{\infty} \frac{\lambda^x}{(x-1)!} \right] - \lambda^2 \\
 &\Leftarrow e^{-\lambda} \left[\sum_{y=0}^{\infty} \frac{\lambda^{y+2}}{y!} + \sum_{z=0}^{\infty} \frac{\lambda^{z+1}}{z!} \right] - \lambda^2 \quad \because \text{put } y = x-2 \\
 &= e^{-\lambda} \left[\lambda^2 \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} + \lambda \sum_{z=0}^{\infty} \frac{\lambda^z}{z!} \right] - \lambda^2 \quad \because \lambda^2 = x-1 \\
 &= e^{-\lambda} [\lambda^2 e^{\lambda} + \lambda e^{\lambda}] - \lambda^2 \quad \because \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} = e^{\lambda} = \sum_{z=0}^{\infty} \frac{\lambda^z}{z!} \\
 &\Leftarrow e^{-\lambda} [e^{\lambda} (\lambda^2 + \lambda)] - \lambda^2 \\
 &= e^{-\lambda} e^{\lambda} (\lambda^2 + \lambda) - \lambda^2 \quad \because e^{\lambda} e^{-\lambda} = 1 \\
 &= \lambda^2 + \lambda - \lambda^2 = \lambda
 \end{aligned}$$

Variance $\sigma^2 = \lambda = \text{Mean of the distribution}$

Standard deviation of P.D is $\sigma = \sqrt{\lambda}$

Recurrence relation for the poisson distribution:

We have by poisson distribution $p(x) = \frac{e^{-\lambda} \lambda^x}{x!}$

$$p(x+1) = \frac{e^{-\lambda} \lambda^{x+1}}{(x+1)!} = \frac{\lambda}{(x+1)} \cdot \frac{e^{-\lambda} \lambda^x}{x!} = \frac{\lambda}{(x+1)} p(x)$$

Thus $p(x+1) = \frac{\lambda}{x+1} p(x)$ (or) $p(x) = \frac{x}{\lambda} p(x-1)$

which is the required relation. With this formula we can find $p(1), p(2), \dots$ if $p(0)$ is given.

→ If the probability that an individual suffers a bad reaction from a certain injection is 0.001, determine the probability that out of 2000 individuals (i) exactly 3 (ii) more than 2 individuals (iii) none (iv) more than one individual suffer a bad reaction.

Sol: Given that $p = \text{prob bad reaction of injection} = 0.001$.

$$n = \text{no. of individuals} = 2000$$

$$\text{Mean } \lambda = np = 2000(0.001) = 2$$

$$[\text{value of } e = 2.718]$$

$$\text{By poisson distribution } p(x) = \frac{e^{-\lambda} \lambda^x}{x!} = \frac{e^{-2} 2^x}{x!} = \frac{2^x}{e^2 x!}$$

$$(i) \text{ Exactly 3 means } p(x=3) = \frac{2^3}{e^2 3!} = \frac{8}{(2.718)^2 (6)} = 0.1804$$

$$(ii) \text{ More than 2 individuals means } p(x > 2) = p(x=3) + \dots$$

$$= 1 - [p(x \leq 2)] = 1 - [p(x=0) + p(x=1) + p(x=2)]$$

$$= 1 - \frac{1}{e^2} \left[\frac{2^0}{0!} + \frac{2^1}{1!} + \frac{2^2}{2!} \right] = 1 - \frac{1}{(2.718)^2} [1+2+2]$$

$$p(x > 2) = 1 - \frac{5}{(2.718)^2} = 1 - 0.67667 = 0.3233$$

$$(iii) \text{ None means } p(x=0) = \frac{2^0}{e^2 (0!)} = \frac{1}{(2.718)^2} = 0.1353$$

$$(iv) \text{ More than one i.e. } p(x > 1) = p(x=2) + p(x=3) + \dots + p(x=2000) = 1 - [p(x \leq 1)]$$

$$= 1 - [p(x=0) + p(x=1)] = 1 - \frac{1}{e^2} \left[\frac{2^0}{0!} + \frac{2^1}{1!} \right]$$

$$p(x > 1) = 1 - \frac{3}{(2.718)^2} [1+2] = 1 - \frac{3}{(2.718)^2} = 1 - 0.496 = 0.594$$

Normal Distribution

Normal distribution is applicable in the following situations:

1. Life of items subjected to wear and tear like tyres, batteries, bulbs, currency notes etc.
2. Length and diameter of certain products like pipes, screws and discs
3. Height and weight of ~~baby~~ at birth.
4. Aggregate marks obtained by students in an examination.
5. Weekly sales of an item in a store.

Def: A random variable x is said to have a Normal distribution if its density function or probability distribution is given by

$$f(x; \mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, -\infty < x < \infty, -\infty < \mu < \infty, \sigma > 0$$

Where μ is the mean and σ is the standard deviation of x . As can be seen, the function called probability density function of the Normal distribution depends on two variables μ and σ .

Mean of Normal distribution:

Consider the Normal distribution with b, σ as the parameters,

$$\text{Then } f(x; b, \sigma) = \frac{e^{-\frac{(x-b)^2}{2\sigma^2}}}{\sigma \sqrt{2\pi}}$$

The mean $M = E(x)$ is given by $\mu = \int_{-\infty}^{\infty} x f(x) dx$.

$$\mu = \frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^{\infty} x e^{-\frac{1}{2} \left(\frac{x-b}{\sigma}\right)^2} dx$$

$$= \frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^{\infty} (\sigma z + b) e^{-\frac{z^2}{2}} dz = \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z e^{-\frac{z^2}{2}} dz + \frac{b}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{z^2}{2}} dz$$

$$= 0 + \frac{b}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z e^{-\frac{z^2}{2}} dz \quad : z e^{-\frac{z^2}{2}} \text{ is odd function and } e^{-\frac{z^2}{2}} \text{ is even function}$$

$$= \frac{2b}{\sqrt{2\pi}} \cdot \frac{\sqrt{\pi}}{\sqrt{2}}$$

$$= \frac{2b\sqrt{\pi}}{2\sqrt{2}} = b.$$

Mean of the Normal distribution $\mu = b$.

Variance of Normal Distribution:

$$\text{By definition Variance } \sigma^2 = E[(x-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx.$$

$$= \frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^{\infty} (x-\mu)^2 e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

$$\text{put } z = \frac{x-\mu}{\sigma}$$

$$= \frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^{\infty} (\sigma z + \mu - \mu)^2 e^{-\frac{z^2}{2}} dz$$

$$dz = \frac{dx}{\sigma}$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (\sigma z)^2 e^{-\frac{z^2}{2}} dz$$

$$= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{-\frac{z^2}{2}} dz \quad \text{Integrand is even function}$$

$$= \frac{2\sigma^2}{\sqrt{2\pi}} \int_0^{\infty} 2t \cdot e^{-t^2} \frac{dt}{\sqrt{2\pi}}$$

$$\text{put } \frac{z^2}{2} = t \Rightarrow dz = \frac{dt}{\sqrt{2\pi}}$$

$$= \frac{2\sigma^2}{2\sqrt{\pi}} \int_0^{\infty} \sqrt{t} \cdot e^{-t} dt$$

$$= \frac{2\sigma^2}{\sqrt{\pi}} \int_0^{\infty} e^{-t} t^{3/2} dt$$

$$= \frac{2\sigma^2}{\sqrt{\pi}} M_{3/2}$$

$$M_n = (n/M_{n-1}) M_{n-1}$$

$$= \frac{2\sigma^2}{\sqrt{\pi}} \left(\frac{3}{2}-1\right) M_{1/2}$$

$$M_{1/2} = \frac{3}{2} M_{-1/2}$$

$$= \frac{2\sigma^2}{\sqrt{\pi}} \cdot \left(\frac{1}{2}\right) \sqrt{\pi}$$

$$M_{1/2} = \sqrt{\pi}$$

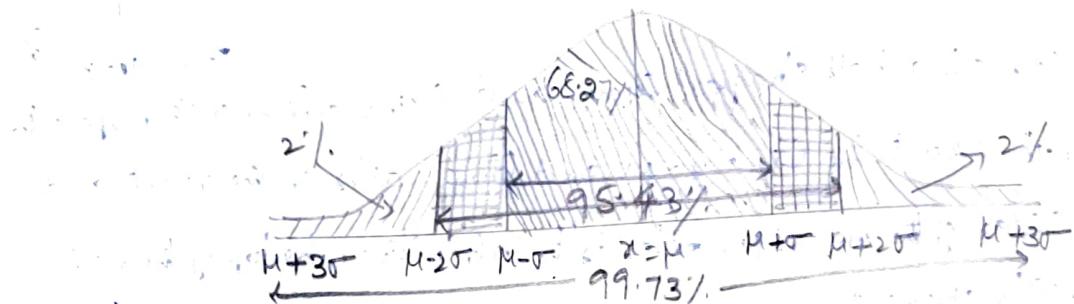
$$\text{Variance} = \sigma^2 = \frac{2\sigma^2}{\sqrt{\pi}} \cdot \left(\frac{1}{2}\right) \sqrt{\pi} = \sigma^2$$

The standard deviation of the Normal distribution = σ .

chief characteristics of the Normal distribution:

1. The graph of the Normal distribution $y = f(x)$ in the xy -plane is known as the normal curve.
 2. The curve is a bell shaped curve and symmetrical with respect to mean i.e. about the line $x = \mu$ and two tails on the right and left sides of the mean (μ) extends to infinity. The top of the bell is directly above the mean μ .
 3. Area under the normal curve represents the total population.
 4. Mean, Median and Mode of the distribution coincide at $x = \mu$ as the distribution is Symmetrical, so normal curve is unimodal.
 5. x -axis is an asymptote to the curve.
 6. Linear combination of independent normal variables is also a normal variate.
 7. The points of inflection of the curve are at $x = \mu \pm \sigma$ and the curve changes from concave to convex at $x = \mu + \sigma$ to $x = \mu - \sigma$.
 8. The probability that the normal variable x with mean μ and standard deviation σ lies between x_1 and x_2 is given by
- $$P(x_1 \leq x \leq x_2) = \frac{1}{\sigma \sqrt{2\pi}} \int_{x_1}^{x_2} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$
- Since ① depends on the two parameters μ and σ , we get different normal curves for different values of μ and σ , it is an impractical task to plot all such normal curves. Instead by putting $z = \frac{x-\mu}{\sigma}$, the R.H.S of equation ① becomes independent of the two parameters μ and σ . Here z is known as the standard variable.
-

9. Area under the normal curve is distributed as follows:



(i) Area of normal curve between $\mu - \sigma$ and $\mu + \sigma$ is 68.27%.
ie. $P(\mu - \sigma < X < \mu + \sigma) = 0.6826.$

(ii) Area of normal curve between $\mu - 2\sigma$ and $\mu + 2\sigma$ is 95.43%.

(iii) Area of normal curve between $\mu - 3\sigma$ and $\mu + 3\sigma$ is 99.73%.

Standard Normal distribution:

The Normal distribution with mean ($\mu = 0$) and

$S.D (\sigma) = 1$ is known as Standard Normal Distribution.

Uses of Normal Distribution:

1. The Normal distribution can be used to approximate Binomial and Poisson distribution.

2. It has extensive use in Sampling theory. It helps us to estimate parameter from statistic and to find confidence limits of the parameter.

3. It has a wide use in testing statistical Hypothesis and Tests of Significance in which it is always assumed that the population from which the samples have been drawn should have normal distribution.

4. It serves as a guiding instrument in the analysis and interpretation of statistical data.

How to find probability Density of Normal Curve:-

The probability that the normal variable X with mean μ and standard deviation σ lies between two specific values x_1 and x_2 with $x_1 \leq x_2$ can be obtained using area under the standard normal curve as follows:

Step 1: perform the change of scale $z = \frac{x-\mu}{\sigma}$ and find z_1 and z_2 corresponding to the values of x_1 and x_2 respectively

Step 2(a): To find $P(x_1 \leq X \leq x_2) = P(z_1 \leq Z \leq z_2)$

Case 1: If both z_1 and z_2 are positive (or both negative) then

$$P(x_1 \leq X \leq x_2) = A(z_2) - A(z_1)$$

= (Area under the normal curve from 0 to z_2)

- (Area under the normal curve from 0 to z_1).



Case 2: If $z_1 < 0$ and $z_2 > 0$ then

$$P(x_1 \leq X \leq x_2) = A(z_2) + A(z_1)$$

Step 2(b): To find $P(Z > z_1)$

Case 1: If $z_1 > 0$ then

$$P(Z > z_1) = 0.5 - A(z_1) \quad [\because P(Z < 0) = P(Z > 0) = \frac{1}{2}]$$



Case 2: If $z_1 < 0$ then $P(Z > z_1) = 0.5 + A(z_1)$.

Step 2(c): To find $P(Z < z_1)$

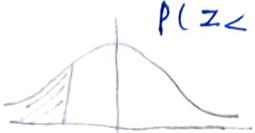
Case 1: If $z_1 > 0$ then $P(Z < z_1) = 1 - P(Z > z_1)$



$$\begin{aligned} P(Z < z_1) &= 1 - P(Z > z_1) = 1 - [0.5 - A(z_1)] \\ &= 0.5 + A(z_1). \end{aligned}$$

Case 2: If $z_1 \leq 0$ then

$$P(Z < z_1) = 1 - P(Z > z_1) = 1 - [0.5 + A(z_1)] = 0.5 - A(z_1).$$



→ For a normally distributed variable with mean 1 and standard deviation 3, find the probabilities that (i) $3.43 \leq x \leq 6.19$
(ii) $-1.43 \leq x \leq 6.19$.

Sol: Given that $\mu = \text{mean} = 1$ and $s.d(\sigma) = 3$.

(i) When $x = 3.43$,

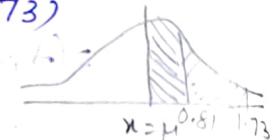
$$Z = \frac{x-\mu}{\sigma} = \frac{3.43-1}{3} = \frac{2.43}{3} = 0.81 = z_1 \text{ (say)}$$

When $x = 6.19$

$$Z = \frac{x-\mu}{\sigma} = \frac{6.19-1}{3} = \frac{5.19}{3} = 1.73 = z_2 \text{ (say)}$$

$$\therefore P(3.43 \leq x \leq 6.19) = P(0.81 \leq Z \leq 1.73)$$

$$= |A(z_2) - A(z_1)|$$



$$= |A(1.73) - A(0.81)| = 0.4582 - 0.2910 \quad \text{From table}$$

$$= 0.1672$$

(ii) When $x = -1.43$

$$Z = \frac{x-\mu}{\sigma} = \frac{-1.43-1}{3} = -0.81 = z_1 \text{ (say)}$$

When $x = 6.19$

$$Z = \frac{x-\mu}{\sigma} = \frac{6.19-1}{3} = 1.73 = z_2 \text{ (say)}$$

$$P(-1.43 \leq x \leq 6.19) = P(-0.81 \leq Z \leq 1.73)$$

$$= A(z_2) + A(z_1)$$

$$= A(1.73) + A(-0.81)$$

$$= A(1.73) + A(0.81) \quad \because A(-z) = A(z)$$

$$= 0.4582 + 0.2910 \quad \text{From table}$$

$$P(-1.43 \leq x \leq 6.19) = 0.7492$$

UNIT-IV:ESTIMATION AND TESTING OF HYPOTHESIS,LARGE SAMPLE TESTS

Lecture Notes

Unit - IV Estimation.

parameters: Quantities appearing in distribution such as p in the binomial distribution and μ and σ in the normal distribution are called parameters.

Estimate: An estimate is a statement made to find an unknown population parameter.

Estimator: The procedure or rule to determine an unknown population parameter is called an Estimator.

Ex: Sample mean is an estimator of population mean because Sample mean is a method of determining the population mean.

A parameter can have one or two or many estimators.

Types of Estimation: There are two types of estimation.

(a) Point Estimation (b) Interval Estimation

(a) Point Estimation: If an estimate of the population parameter is given by a single value, then the estimate is called a point estimation of the parameter.

(b) Interval Estimation: If an estimate of the population parameter is given by two different values between which the parameter may be considered to lie, then the estimate is called an interval estimation of the parameter.

Ex: If the height of a student is measured as 162 cm, then the measurement gives a point estimation.

But if the height is given as (163 ± 3.5) cms, then the height lies between 159.5 cm

and 166.5 cm and the measurement gives an interval estimation.

Properties of Estimation: An estimator is not expected to estimate the population parameter without error. An estimator should be close to the true value of unknown parameter.

Unbiased and Biased Estimates: A statistic is said to be an unbiased estimator of the corresponding parameter if the mean of the sampling distribution of the statistic is equal to the corresponding population parameter. Otherwise the statistic is called a Biased estimator of the corresponding parameter. The values of statistics in the above two cases are called unbiased and biased estimates respectively.

Let t be a statistic and θ be the corresponding parameter and $E(t) = \theta$, then t is an unbiased estimator of θ . Otherwise t is a biased estimator of θ and the bias is $E(t) - \theta$.

Unbiased Estimator: A statistic or point estimator $\hat{\theta}$ is said to be an unbiased estimator of the parameter θ if $E(\hat{\theta}) = \theta$.

Theorem: Shows that Sample mean \bar{x} is an unbiased estimator of population mean μ i.e. $E(\bar{x}) = \mu$.

Proof: Let x_1, x_2, \dots, x_n be a random sample drawn from a given population with mean μ and variance σ^2 .

$$\begin{aligned} \text{Then } E(\bar{x}) &= E\left(\frac{1}{n} \sum_{i=1}^n x_i\right) = \frac{1}{n} E(x_1 + x_2 + \dots + x_n) \\ &= \frac{1}{n} [E(x_1) + E(x_2) + \dots + E(x_n)], \\ &= \frac{1}{n} [M + M + \dots + M] \quad (E(x_i) = \mu \text{ for all } i) \\ &= \frac{nM}{n} = \mu. \end{aligned}$$

Hence the sample mean \bar{x} is an Unbiased estimator of the population mean μ .

Q1: Show that s^2 is an unbiased estimator of the parameter σ^2 .

Sol: Let us write $\sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n [(x_i - \mu) - (\bar{x} - \mu)]^2$

$$= \sum_{i=1}^n (x_i - \mu)^2 - 2(\bar{x} - \mu) \sum_{i=1}^n (x_i - \mu) + n(\bar{x} - \mu)^2.$$

$$= \sum_{i=1}^n (x_i - \mu)^2 - n(\bar{x} - \mu)^2$$

Now $E(s^2) = E\left[\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}\right] = \frac{1}{n-1} \left[\sum_{i=1}^n E(x_i - \mu)^2 - nE(\bar{x} - \mu)^2 \right]$

$$= \frac{1}{n-1} \left[\sum_{i=1}^n \sigma_{x_i}^2 - n\sigma_{\bar{x}}^2 \right]$$

However $\sigma_{x_i}^2 = \sigma^2$, for $i=1, 2, \dots, n$ and $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$.

$$\therefore E(s^2) = \frac{1}{n-1} \left[n\sigma^2 - n\frac{\sigma^2}{n} \right] = \frac{\sigma^2(n-1)}{(n-1)} = \sigma^2.$$

Q2: Although s^2 is an unbiased estimator of σ^2 , s^2 on the other hand is a biased estimator of σ with the bias becoming insignificant for large samples. This example illustrates why we divide by $(n-1)$ rather than n when the variance is estimated.

Q3: Prove that for a random sample of size n , x_1, x_2, \dots, x_n taken from an infinite population, $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$ is an unbiased estimator of the parameter σ^2 but $\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$ is not.

Proof: Let μ be the population mean. Then $E(x_i) = \mu$ and

$$\text{Var}(x_i) = E[(x_i - \mu)^2] = \sigma^2, \quad i=1, 2, \dots, n.$$

If S be the sample SD then $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \rightarrow \textcircled{O}$

Now $s^2 = \frac{1}{n} \sum_{i=1}^n [x_i^2 - 2x_i \bar{x} + \bar{x}^2] \rightarrow \textcircled{O}$

$$= \frac{1}{n} \sum_{i=1}^n x_i^2 - 2\bar{x} \sum_{i=1}^n \frac{x_i}{n} + \frac{1}{n} \sum \bar{x}^2.$$

$$= \frac{\sum_{i=1}^n x_i^2}{n} + 2\bar{x}^2 + \bar{x}^2 = \frac{\sum_{i=1}^n x_i^2}{n} + \bar{x}^2$$

$$= \frac{\sum_{i=1}^n (x_i - \mu)^2}{n} + (\bar{x} - \mu)^2.$$

$$\left[\therefore \frac{\sum (x_i - \mu)^2}{n} + (\bar{x} - \mu)^2 = \frac{\sum x_i^2}{n} - 2\mu \frac{\sum x_i}{n} + n \cdot \frac{\mu^2}{n} - (\bar{x}^2 + \mu^2 - 2\bar{x}\mu) \right]$$

$$\therefore E(s^2) = E \left[\frac{\sum (x_i - \mu)^2}{n} + (\bar{x} - \mu)^2 \right] = E \left[\frac{\sum (x_i - \mu)^2}{n} \right] + E(\bar{x} - \mu)^2$$

$$= E \frac{\sum (x_i - \mu)^2}{n} - E(\bar{x} - \mu)^2 = \frac{\sum \text{var}(x_i)}{n} - \text{var}(\bar{x})$$

$$= \frac{\sum \sigma^2}{n} - \frac{\sigma^2}{n} \quad (\because \bar{x} = \frac{\sum x_i}{n})$$

$$= \frac{n\sigma^2}{n} - \frac{\sigma^2}{n} = \sigma^2(1 - \frac{1}{n}) = \frac{(n-1)\sigma^2}{n}$$

Thus $E(s^2) \neq \sigma^2$.

Hence s^2 is a biased estimator of σ^2 .

Let us consider $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{n}{n-1} s^2$ (by G.)

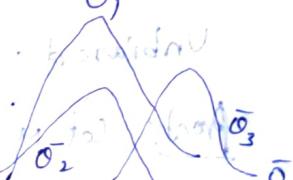
Then $E(s^2) = E \left[\frac{n}{n-1} s^2 \right] = \frac{n}{n-1} E(s^2) = \frac{n}{n-1} \times \frac{n-1}{n} \sigma^2$

Hence $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$ is an unbiased estimator of σ^2 .

Most Efficient Estimator: If we consider all possible unbiased estimators of some parameter θ , the one with the smallest variance is called the Most efficient estimator.

In the fig, we illustrate the sampling distribution of three different estimators $\bar{\theta}_1$ and $\bar{\theta}_2$ and $\bar{\theta}_3$ all estimating θ .

It is clear that only $\bar{\theta}_1$ and $\bar{\theta}_2$ are unbiased, since their distributions are centered at θ . The estimator $\bar{\theta}_1$ has a smaller variance than $\bar{\theta}_2$ and is therefore more efficient. Hence our choice for an estimator of θ among the three considered would be $\bar{\theta}_1$.



Properties of Estimators: A Good estimator is one which is as close to the true value of the parameter as possible. The important properties of a good estimator are:

(i) Consistency (ii) Unbiasedness (iii) Efficiency and (iv) Sufficiency.

(i) An estimator $\bar{\theta}_n$ of a parameter θ is consistent if it converges to θ ,

(ii) A statistic $\bar{\theta}$ is said to be unbiased estimator of θ if $E(\bar{\theta}) = \theta, \forall \theta$.

(iii) A statistic $\bar{\theta}_1$ is said to be a more efficient unbiased estimator of the parameter θ than the statistic $\bar{\theta}_2$ if

(a) $\bar{\theta}_1$ and $\bar{\theta}_2$ are both unbiased estimators of θ .

(b) $V(\bar{\theta}_1) < V(\bar{\theta}_2)$

(iv) An estimator is said to be sufficient for a parameter, if it contains all the information in the sample regarding the parameter.

Formulas:

- (1) Standard error of $\bar{x} = \frac{\sigma}{\sqrt{n}}$, Maximum error $E = z_{\alpha/2} \left(\frac{\sigma}{\sqrt{n}} \right)$.

- (2) Confidence limits for \bar{x} is $[\bar{x} - z_{\alpha/2} (S.E. \bar{x}), \bar{x} + z_{\alpha/2} (S.E. \bar{x})]$

- (3) Sample Size $n = \left(\frac{z_{\alpha/2} \sigma}{E} \right)^2$. $z_{\alpha/2}$ is table value.

- (4) Confidence interval for μ, σ Known is

$$\bar{x} - z_{\alpha/2} \left(\frac{\sigma}{\sqrt{n}} \right) < \mu < \bar{x} + z_{\alpha/2} \left(\frac{\sigma}{\sqrt{n}} \right).$$

Where $z_{\alpha/2}$ is the Z-value leaving an area of $\alpha/2$ to its right.

Bayesian Estimation:- Combining the prior beliefs about the possible values of μ with direct sample evidence the posterior distribution of μ in Bayesian estimation is approximated by normal distribution with

$$\mu_1 = \frac{n\bar{x}\sigma_0^2 + \mu_0\sigma^2}{n\sigma_0^2 + \sigma^2} \text{ and } \sigma_1 = \sqrt{\left[\frac{\sigma_0^2}{n\sigma_0^2 + \sigma^2} \right]}$$

Where n = Sample Size, \bar{x} = Sample mean and S is the Standard Deviation of Sample. Use $S = \sigma$.

Here μ_1 and σ_1 are known as the mean and S.D. of the posterior distribution. In the Computation of μ_1 and σ_1 , σ^2 is assumed to be known, when σ^2 is unknown, which is generally the case, is replaced by Sample Variance s^2 provided $n \geq 30$ (Large Sample).

Bayesian interval for μ :

(1- α) 100% Bayesian interval for μ is given by

$$\mu_1 - z_{\alpha/2} \sigma_1 < \mu < \mu_1 + z_{\alpha/2} \sigma_1$$

→ In a study of an automobile insurance a random sample of 80 body repair costs had a mean of £ 472.36 and the S.D of £ 62.35. If used as a point estimate to the true average repair costs, with what confidence we can assert that the maximum error doesn't exceed £ 10.

Sol: Size of the Random Sample $n = 80$, Mean $\bar{x} = 472.36$.

S.D, $\sigma = 62.35$, maximum error of estimate $E \leq 10$.

$$\text{We have } E_{\max} = Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \Rightarrow Z_{\alpha/2} = \frac{E_{\max} \sqrt{n}}{\sigma} = \frac{10 \sqrt{80}}{62.35} = 1.4345$$

$$\therefore Z_{\alpha/2} = 1.43. \quad \text{The area when } Z = 1.43 \text{ from table is } 0.4236.$$

$$\therefore \alpha/2 = 0.4236 \Rightarrow \alpha = (0.4236)^2 = 0.17872.$$

∴ Confidence $= (1-\alpha) 100\% = 84.72\%$.

Hence we are 84.72% confidence that the maximum error is £ 10.

→ What is the size of the smallest sample required to estimate an unknown proportion to within a maximum error of 0.06 with at least 95% confidence.

Sol: Given maximum error $E = 0.06$.

Confidence limit $= 95\% \Rightarrow (1-\alpha) 100 = 95$.

$$\Rightarrow 1-\alpha = 0.95 \Rightarrow \alpha = 0.05 \Rightarrow \alpha/2 = 0.025$$

$$\therefore Z_{\alpha/2} = 1.96.$$

Here p is not given, so we take $P = \frac{1}{2}$, $Q = \frac{1}{2}$ and $Z_{\alpha/2} = 1.96$.

$$\text{Hence } n = \left[\frac{Z_{\alpha/2}}{E} \right]^2 P Q = \left(\frac{1.96}{0.06} \right)^2 \frac{1}{2} \cdot \frac{1}{2} = 266.78 \text{ (Table value)}$$

$$n = \frac{1}{4} \left[\frac{1.96}{0.06} \right]^2 = 266.78.$$

∴ Sample size $n = 267$.

Testing of Hypothesis (Large Sample Test).

Population: Population or Universe is the aggregate or totality of

Statistical data forming a subject of investigation.

Ex: The population of the heights of Indians.

The population of Nationalised Banks in India etc.

Size of the population: The no. of observations in the population is defined to be the size of the population. It may be finite or infinite. It is denoted by N .

Sample: A portion of the population which is examined with a view to determining the population characteristics is called a Sample.

No. of objects in the Sample is called Sample Size, denoted by n .

Sample are two types:

i. Large Sample: If the size of the sample $n \geq 30$ the sample is

Called Large Sample.

ii. Small Sample: If the size of the sample $n < 30$ the sample is

Called Small Sample or Exact Sample.

Test of Hypothesis: We need to decide whether to accept or reject a statement about the parameter. This statement is called a hypothesis and the decision-making procedure about the hypothesis is called Hypothesis testing.

Procedure For Testing of Hypothesis

Step 1: Null Hypothesis: Define or set up a Null Hypothesis H_0 , taking into consideration the nature of the problem and data involved.

Step 2: Alternative Hypothesis: Set up the alternative hypothesis H_1 , so that we could decide whether we should use one-tailed or two-tailed test.

Step 3: Level of Significance (L.O.S): Select the appropriate level of significance (α) depending on the reliability of the estimate and permissible risk. That is a suitable α is selected in advance if it is not given in the problem.

Step 4: Test statistic: Compute the test statistic $Z = \frac{t - E(t)}{S.E.g(t)}$ under the null hypothesis. Here t is a sample statistic and $S.E$ is the S.D. of t .

Step 5: Conclusion: we compare the computed value of the test statistic Z with the critical value Z_α at given level of significance.

If $|Z| < Z_\alpha$, we conclude that it is not significant. we accept the Null Hypothesis.

If $|Z| > Z_\alpha$, then the difference is significant and hence the hypothesis of Null is rejected at the level of significance α .

It is important to note that in case of one-tail test, significance depends on the sign of the deviation from the null hypothesis.

That is if the observed value is greater than the critical value, then it is significant.

Null Hypothesis: A null hypothesis is the hypothesis which asserts that there is no significant difference between the salutistic and the population parameter and whatever observed difference is there is merely due to fluctuations in sampling from the sample population. It is denoted by H_0 .

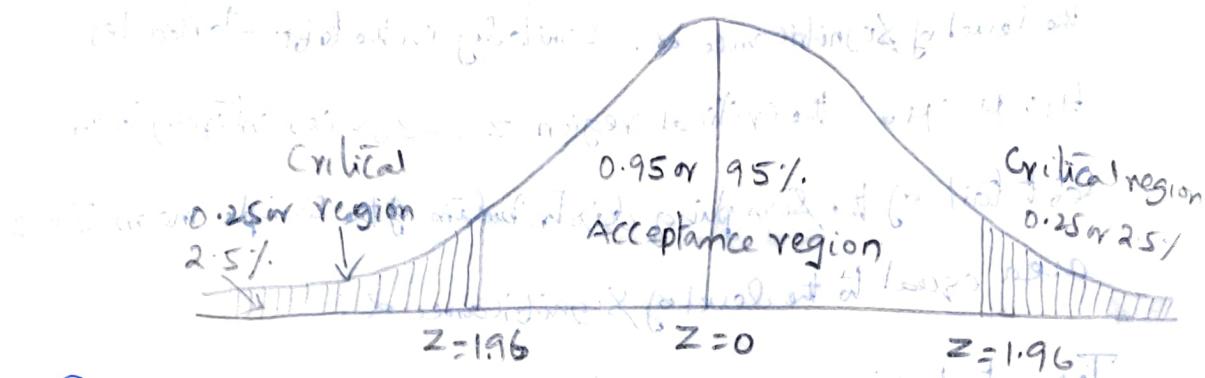
Alternative Hypothesis: Any hypothesis which contradicts the Null Hypothesis is called an Alternative Hypothesis. It is denoted by H_1 . The two hypotheses H_0 and H_1 are such that if one is true, the other is false and vice versa.

Level of Significance: The level of significance denoted by α is the confidence with which we rejects or accepts the Null hypothesis i.e. it is the maximum possible probability with which we are willing to risk an error in rejecting H_0 when it is true.

Errors of Sampling: There are two types of errors:

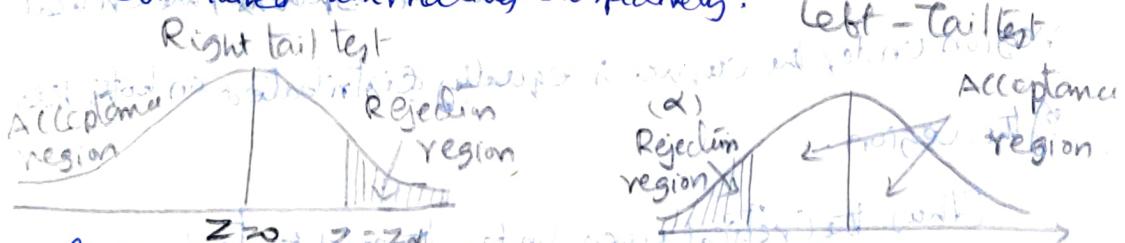
- Type I error: Reject H_0 when it is true.
If the Null Hypothesis H_0 is true but it is rejected by test procedure, then the error made is called Type I error or α error.
- Type II error: Accept H_0 when it is wrong i.e. accept H_0 when H_1 is true. If the Null Hypothesis is false but it is accepted by test procedure, then error is called Type II error or β error.

Critical Region or Rejection Region: A region corresponding to a statistic t in the sample space S which leads to the rejection of H_0 is called Critical Region or Rejection Region. Those regions which lead to acceptance of H_0 give us a region called Acceptance Region.



One-Tailed Test: If we have to test whether the population μ has a specified value μ_0 , then the Null Hypothesis is $H_0: \mu = \mu_0$ and the alternative hypothesis may be (i) $H_1: \mu > \mu_0$ or (ii) $H_1: \mu < \mu_0$.

The alternative hypothesis in (i) and (ii) are known as right-tailed and left-tailed alternatives respectively.



If the alternative hypothesis H_1 in a test of a statistical hypothesis be one-tailed (ie either right-tailed or left-tailed but not both) then the test is called a one-tailed test. For example to test whether the population mean $\mu = \mu_0$ we have $H_0: \mu = \mu_0$ against the alternative hypothesis H_1 given by

(i) $H_1: \mu > \mu_0$ (right-tailed) or (ii) $H_1: \mu < \mu_0$ (left-tailed)

and the corresponding test is a single-tailed or one-tailed.

In the right-tail test $H_1: \mu > \mu_0$, the critical region (or rejection region) $z > z_0$ lies entirely in the right tail.

Sampling distribution of sample mean \bar{x} with area equal to the level of significance α . Similarly in the left-tailed test ($H_1: \mu < \mu_0$) the critical region $z < -z_0$ lies entirely in the left tail of the sampling distribution of the sample mean \bar{x} with area equal to the level of significance α .

Two-tailed Test: Suppose we want to test the Null Hypothesis

$H_0: \mu = \mu_0$ against the Alternative Hypothesis $H_1: \mu \neq \mu_0$.

Since H_1 is two-tailed, alternative hypothesis, the critical region under the curve is equally distributed on both sides of the region.

Thus the critical area under the right-tail

= The critical area under the left-tail

= Half of the total area.

= $\frac{1}{2}$ Probability of rejection.

= $\frac{\alpha}{2}$ with critical statistic $z_{\alpha/2}$

Where α is the level of significance

The critical region is then $z \leq -z_{\alpha/2}$ or $z_{\alpha/2} \leq z$.

Critical values of z for both two-tailed and one-tailed tests at 1%, 5% and 10% level of significance are given in the table below.

Critical values of z .

Level of Significance 1%, 5%, 10%

Critical values for two tailed test $|z_{\alpha/2}| = 2.58$, $|z_{\alpha/2}| = 1.96$, $|z_{\alpha/2}| = 1.645$.

Critical value for Right tailed test $z_{\alpha} = 2.33$, $z_{\alpha} = 1.645$, $z_{\alpha} = 1.28$.

Critical value for left tailed test $z_{\alpha} = -2.33$, $z_{\alpha} = -1.645$, $z_{\alpha} = -1.28$.

Under the large samples tests we have the following types:

1. Testing of Significance for Single mean.
2. Testing of Significance for difference of means.
3. Testing of Significance for Single proportion.
4. Testing of Significance for difference of proportions.

Working rule for Testing of Significance for Single mean:

1. Null Hypothesis: $H_0: \bar{x} = \mu$ i.e. there is no significance difference between the sample mean and population mean.
2. Alternative Hypothesis $H_1: \bar{x} \neq \mu$, or $\bar{x} > \mu$ or $\bar{x} < \mu$.
3. Level of Significance: set the level of significance α .
4. Test statistic. We have the following formula

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

Where μ is population mean, \bar{x} is sample mean.

5. Decision: finding the critical value $Z_{\alpha/2}$ of Z at level of significance α from the normal table.

- (a) If $|z| \leq Z_{\alpha/2}$ we accept the Null Hypothesis H_0 .
- (b) If $|z| > Z_{\alpha/2}$, we reject the Null Hypothesis H_0 .

→ According to the norms established for a mechanical aptitude test, persons who are 18 years old have an average height of 73.2 with a standard deviation of 8.6. If 4 randomly selected persons of that age average 76.7, test the hypothesis $H_0: \mu = 73.2$ against the alternative hypothesis $H_1: \mu > 73.2$ at the 0.01 level of significance.

Sol: Given $n = 4$, $\mu = 73.2$, \bar{x} = mean of the sample = 76.7,

and $\sigma = \text{S.D. of Population} = 8.6$.

- ① Null Hypothesis: $H_0: \mu = 73.2$. (one-tailed test)
- ② Alternative Hypothesis: $H_1: \mu > 73.2$. (Right-tailed test)
- ③ Level of Significance: $\alpha = 99\%$. (or Probability 0.01)

Table value for Z at 99%. Look in $Z_{\alpha} = 2.33$.

$$\textcircled{4} \text{ Test statistic: } z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{76.7 - 73.2}{8.6/\sqrt{4}} = 0.814.$$

\therefore Calculated $Z = 0.814$ < Table value of z_{α} .

- ⑤ Inference: Hence calculated $Z <$ Table value of z_{α} .
∴ The Null Hypothesis H_0 is accepted. That is \bar{x} and μ don't differ significantly.

UNIT-V:SMALL SAMPLE TESTS

Lecture Notes

5. Test of Significance (Small Sample Test)

Small Sample Test: If the size of the sample, ($n < 30$) less than 30 we say the Small Sample Test

Degree of Freedom (d.f.):

The number of independent variates which make up the statistic is known as the degree of freedom (d.f.) and it is denoted by ν (the letter 'Nu' of the Greek alphabet). In other words it is the number of values in a set of data which may be assigned arbitrarily (or) it refers to the number of "independent constraints" in a set of data.

1. t-distribution (or) Student's t-distribution:

If $\{x_1, x_2, \dots, x_n\}$ be any random sample of size n drawn from a normal population with mean μ and variance σ^2 , then the test statistic t is defined by $t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$,

Where \bar{x} = Sample mean

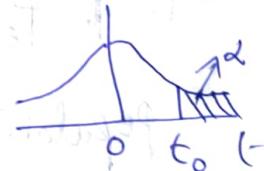
$$S^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} \text{ is an unbiased estimate of } \sigma^2$$

μ = population mean, n is sample size.

Properties of t-distribution :-

1. The shape of t-distribution is bell-shaped

which is similar to that of a normal distribution
and is symmetrical about mean



2. The t-distribution curve is also asymptotic to the t-axis, i.e. the two tails of the curve on both sides of $t=0$ extends to infinity.

3. It is symmetrical about the line $t=0$.

4. The form of the probability curve varies with degrees of freedom i.e. with sample size.

5. It is Unimodal with mean = Median = Mode.

6. The mean of standard normal distribution and as well as t-distribution is zero but the

Variance of t-distribution depends upon the parameter v which is called the degree of freedom.

7. The variance of t-distribution exceeds 1, but approaches 1 as $n \rightarrow \infty$.

Applications of the t-distribution:

The t-distribution has a wide number of applications in statistics. Some of ~~not~~ of them are given below.

1. To test the significance of the sample mean, when population variance is not given.
2. To test the significance of the mean of the sample i.e. to test if the sample mean differs significantly from the population mean.
3. To test the significance of the difference between two sample means or to compare two samples.
4. To test the significance of an observed sample Correlation Coefficient and sample Regression Coefficient.

2. Chi-Square (χ^2) Distribution:

Chi-Squared distribution is a continuous probability distribution of a continuous random variable x with probability density function

Given by $f(x) = \begin{cases} \frac{1}{2^{v/2} \Gamma(v/2)} x^{(v/2)-1} e^{-x/2}, & x > 0 \\ 0, & \text{otherwise.} \end{cases}$

Where v is a positive integer is the only single parameter of the distribution, also known as "degrees of freedom" (d.f.)

Properties of χ^2 -distribution:

1. χ^2 -distribution Curve is not symmetrical, lies entirely in the first quadrant and hence not a normal curve, since χ^2 varies from zero.
2. It depends only on the degrees of freedom v .
3. If χ_1^2 and χ_2^2 are two-distribution both are independent with v_1 and v_2 degrees of freedom, the $\chi_1^2 + \chi_2^2$ will be chi-Squared.

distribution with $(v_1 + v_2)$ degrees of freedom.

4. Here α denotes the area under the Chi-Square distribution to the right of χ^2_{α} , so χ^2_{α} represents the χ^2 -value such that the area under the Chi-Square curve to its right is equal to α .

5. Mean = v and Variance = $2v$.

Applications of χ^2 -distribution :-

1. To test the goodness of fit.
2. To test the independence of attributes.
3. To test the homogeneity of independent estimators of the population variance.
4. To test the homogeneity of independent estimators of the population Correlation Coefficient.

F-distribution : F-distribution is defined by

$$F = \frac{\text{Greater variance}}{\text{Smaller variance}} = \frac{s_1^2}{s_2^2}$$

Which follows F-distribution with $v_1 = n_1 - 1$ and $v_2 = n_2 - 1$ degrees of freedom.

Test of Significance For Small Samples.

→ A Sample of 26 bulbs gives a mean life of 990 hours with a S.D of 20 hours. The manufacturer claims that the mean life of bulbs is 1000 hours. Is the sample not upto the standard.

Sol: Here Sample Size $n = 26 < 30$ (Small sample)

Sample mean $\bar{x} = 990$.

Population mean $\mu = 1000$ and $S.D = 20 = S$ (say)
Degrees of freedom $= n - 1 = 26 - 1 = 25$.

1. Null Hypothesis H_0 : The Sample is upto the standard

2. Alternative Hypothesis H_1 : $\mu < 1000$ (left-tailed test)

3. Level of Significance: $\alpha = 0.05$

At $\alpha = 0.05$ i.e. 95% Confidence 25 degree of freedom for left tailed test is 1.708.

4. Test statistic $t = \frac{\bar{x} - \mu}{S/\sqrt{n-1}} = \frac{990 - 1000}{20/\sqrt{25}} = -2.5$

5. Conclusion: Calculated $t >$ table t , we reject the H_0 . We conclude that the sample is not upto the standard.

Test procedures for two students' t-test
for difference of means:

$$H_0: \mu_1 = \mu_2$$

$$H_A: \mu_1 \neq \mu_2$$

LOS: Same

$$\text{Test Stat. } t = \frac{\bar{x}_1 - \bar{x}_2}{s}$$

$$s = \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$\text{Where } s^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$$