Course Code: BEEF183T10	MATHEMATICS III (Probability and Statistics)	L	T	P	Credit	CIA Marks **40**
Course Category: BSC		3	1	0	4	SEE Marks **60**

COURSE OBJECTIVES:

The objective of this course is to familiarize the students with statistical techniques. It aims to equip the students with standard concepts and tools at an intermediate to advanced level that will serve them well towards tackling various problems in the discipline.

COURSE OUTCOMES:

Students who successfully complete this course should be able to demonstrate and understanding of:

Cos No.	Course Outcomes	Bloom's level
CO1	Basic probability axioms and rules and the moments of discrete and continuous random variables as well as be familiar with common named discrete and continuous random variables.	Remembering, Understanding Applying
CO2	How to derive the probability function of transformations of random variables and use these techniques to generate data from various distributions.	Remembering, Understanding Remembering, Understanding Applying
CO3	How to calculate and apply measures of location and measures of dispersion in grouped and ungrouped data cases.	Applying
CO4	Test of Hypothesis as well as calculate confidence interval for a population parameter for single sample and two sample cases.	Analyzing, Evaluating
CO5	How to translate real-world problems into probability models. Also how to collect data, analyze and deduce information from a real time survey without any unwilling bias	Evaluating , create

Mapping of Course Outcome to Program Outcomes:

Course Outcomes	Program Outcomes											
	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12
CO1	S	S	S	S	M	M	M	-	M	M	M	M
CO2	S	S	S	S	M	M	M	-	M	M	M	M
CO3	S	S	S	S	M	M	M	-	M	M	M	L
CO4	S	S	S	S	M	M	M	-	M	M	M	L
CO5	S	S	S	S	M	M	M	-	M	M	M	L

SYLLABUS:

MODULEI:

BASIC PROBABILITY:

Probability spaces, conditional probability, Independent random variables, sums of independent random variables, Bayes' Theorem, Discrete and Continuous one dimensional random variables - Expectations, Moments, Variance of a sum, Moment generating function, Tchebyshev's Inequality.

MODULE II

PROBABILITY DISTRIBUTIONS:

Discrete Distributions – Binomial, Poisson and Negative Binomial distributions, Continuous Distributions - Normal, Exponential and Gamma distributions.

MODULE III

BASIC STATISTICS:

Measures of Central tendency: Averages, mean, median, mode, Measures of dispersion – Range, Mean deviation, Quartile deviation and Standard deviation, Moments, skewness and Kurtosis, Correlation and regression – Rank correlation.

MODULE IV

APPLIED STATISTICS:

Curve fitting by the method of least squares- fitting of straight lines, second degree parabolas and more general curves. Test of significance: Large sample test for single proportion, difference of proportions, single mean, difference of means, and difference of standard deviations.

MODULE V

SMALL SAMPLES:

Test for single mean, difference of means and correlation coefficients, test for ratio of variances - Chi-square test for goodness of fit and independence of attributes.

LEARNING RESOURCES:

Text Books:

1. T. Veerarajan, Probability, Statistics and Random Processes, Third edition, Tata McGraw-Hill, New Delhi, 2010.
2. S.P. Gupta, Statistical Methods, 31st edition, Sultan chand and sons, New Delhi, 28th Edition 2002.

References:

1. Erwin Kreyszig, Advanced Engineering Mathematics, 9th Edition, John Wiley & Sons,10th Edition Dec 2010.
2. B.S. Grewal, Higher Engineering Mathematics, Khanna Publishers, 43th Edition 2000.
3. S. Ross, A First Course in Probability, 6th Ed., Pearson Education India, 9th Edition 2013.
4. W. Feller, An Introduction to Probability Theory and its Applications, Vol. 1, 3rd Ed.,Wiley, 1968.
5. N.P. Bali and Manish Goyal, A text book of Engineering Mathematics, LaxmiPublications, Reprint, 9th Edition 2015.

Online resources:

1. www.nptl.co.in
2. www.electrical4u.com

ASSESSMENT PATTERN:

Bloom's Category	**Continuous Assessment Tests**		**Assignment (10)**	**Terminal Examination (60)**
	(15)	**(15)**		
Remember	7	0	10	20
Understand	8	-		15
Apply	0	8		10
Analysis	0	4		10
Evaluate	0	3		5
Create	0	0		0
Total	**15**	**15**		**60**
	A	**B**	**C**	**D**

A+B+C+D = 100

UNIT-I

Basic Probability

Random Experiment

If an experiment is repeated under the same conditions, any number of times, it does not give unique results but may result in any of the several possible outcomes

Thus an experiment whose outcome cannot be predicted is called a random experiment or trial and the outcomes are known as events or cases.

Sample space

The set of all possible outcomes of a random experiment is called a sample space and is denoted by S.

Favourable Events

The number of outcomes favourable to an event in an experiment is the number of outcomes which entail the happening of the event.

Example

In tossing two coins the cases favourable to the event of getting a head are HT, TH, HH

Mutually Exclusive events

Two events A and B are said to be mutually exclusive if they can not occur simultaneously.

Note

If A and B represent mutually exclusive events then they are disjoint, that is $A \cap B = \phi$ where ϕ is the null set.

Example

1. when we toss a coin, either head or tail can be up, but both cannot be up at a time.

2. when we throw a dice the outcomes getting 1, 2, 3 . . . . 6 are mutually exclusive events.

Hint : Mutually exclusive events are applicable for a single trail only.

If a coin is tossed twice, the head appearing in the first trail will not affect the appearing of the tail in the next trail.

Equally like events

The events are said to be equally likely if none of them is expected to occur in preference to the other.i.e each one of them has an equal chance of happening.

Example

In tossing of a coin, getting a head and tail are equally likely events.

Exhaustive Events

Outcomes are said to be Exhaustive when they include all possible out comes.

Example

In drawing two cards from a pack of 52 cards, the exhaustive number of cases in $^{52}c_2$

In the case of throwing two dice the exhaustive number of cases is 36 ($=6^2$).

Independent Event

Two or more events are considered to be independent if the occurrence or non-occurrence of an event does not affect the occurrence or non occurrence of the other.

Example

In successive tossing of a coin, the event of getting a head or tail in the first toss does not affect the event of getting a head or tail in the second toss.

Dependent Events

The events are said to be dependent if the occurrence or non-occurrence of one event in any trial affects the occurrence of other events in other trails.

Mathematical classical or a priority definition of probability

If a trail results in n exhaustive mutually exclusively and equally likely cases and m of them are favourable to the happening of the event A, then the probability of happening of A is given by

$$P(A) = p = \frac{\text{number of favaourable cases}}{\text{Total number of exhaustive cases}} = \frac{\text{n(A)}}{\text{n(S)}} = \frac{\text{m}}{\text{n}}$$

For example

In throwing a dice, the possible cases are

If A is an event of getting a number 5, then n(A) =1

The probability of getting 5 is

Note :

i) The probability p of the happening of an event is also known as the probability of success.

ii) The probability q=l-p of the non-happening of the event is known as the probability of failure.

iii) If P(A)=1, then A is called a certain event.

iv) If P(A)=0, A is called an impossible event.

v) If the exhaustive number of cases in a trial is infinite, then this definition of classical probability breaks down.

vi) If the events are not equally likely then this definition of mathematical probability breaks down.

Statistical or a Post Priori or Emperical definition of probability

If a trial is repeated n times under essentially homogeneous and identical conditions and let an event A occur m times out of n trials, n becomes indefinitely large then the probability p of the happening of A is given by $\mathrm{P(A)} = p = \lim_{n\to\infty} \frac{m}{n}$

Axomatic Definition of Probability

Let S be the sample space and A be an event associated with a random experiment. Then the probability of the event A, denoted by P(A) is a real number satisfying the following axioms

(i) $0 \le P(A) \le 1$

(ii) $\mathrm{P}(S) = 1, \mathrm{P}(\phi) = 0$

(iii) **Addition theorem**

$\mathrm{If\ A_1, A_2, ..., A_n}$ are mutually exclusive events then

$\mathrm{P(A_1 \cup A_2 \cup .. \cup .A_n) = P(A_1) + P(A_2) + ... + P(A_n)}$

i.e., $\mathrm{P(\cup A_n) = \sum P(A_n)}$

Standard Results on Probability

(i) The probability of an impossible event is zero. i.e., $\mathrm{P}(\phi) = 0$

(ii) $\mathrm{P}(\overline{\mathrm{A}}) = 1 - \mathrm{P(A)}$

(iii) $\mathrm{If\ B} \subset \mathrm{A\ thn\ P(B)} \le P(A)$

(iv) If A and B are two events which are not disjoint then

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

This is known as additive law of probability

(v) If A,B,C are any three events then
$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(C \cap A) + P(A \cap B \cap C)$

(vi) Two events A and B are independent if $P(A \cap B) = P(A)P(B)$

This is known as **multiplication law of probability**

(vii) If A and B are independent events then $\overline{\mathrm{A}}$ and $\overline{\mathrm{B}}$ also independent

Examples

1. In tossing a coin, find the probability of getting a head?

Solution

$$P(\text{getting a head}) = \frac{\text{Number of favourable events}}{\text{Number of exhaustive events}} = \frac{1}{2}.$$

2. From a bag containing 10 black and 12 white balls, a ball is drawn at random. What is the probability that it is black?

Solution

Let A be the event of selecting a black ball .

Total number of balls = 10+12=22

Total number of possible (exhaustive) ways of choosing one ball = $22c_1 = 22$ ways

n(S)=22

Out of 10 black balls, the number of ways of choosing one black ball = $10c_1 = 10$

n(A)=10

$\therefore$

Probability of getting a black ball = $\frac{n(A)}{n(S)} = \frac{10}{22} = 0.4545$.

3. If at least one child in a family of three children is a boy, what is the probability that all three are boys?

Solution

The sample space is

S= {BBB,BBG,BGB,GBB,GGB,GBG,BGG} where B represents boy and G represents a girl.

n(S)=7

P(all three are boys)= $\frac{n(A)}{n(S)} = \frac{1}{7}$

4. A problem is given to 3 students A, B, C whose chances of solving it are 1/2, 1/3, 1/4 respectively. what is the probability that

i) The problem is solved?

ii) Exactly one of them solves the problem.

Solution

Let A, B, C be the events that A, B, C respectively solve the problem.

P(A)=1/2, P(B)=1/3, P(C)=1/4

$P(\overline{A}) = 1/2, \mathrm{P}(\overline{\mathrm{B}}) = 2/3, P(\overline{C}) = 3/4$.

P(The problem is not solved)= $P(\overline{A} \cap \overline{B} \cap \overline{C})$

$$= P(\overline{A})P(\overline{B})P(\overline{C})$$
$$= \frac{1}{2} \times \frac{2}{3} \times \frac{3}{4}$$
$$= \frac{1}{4}$$

Therefore, P(the problem is solved)=1-1/4=3/4.

ii) P(Exactly one of them solves the problem)= $P\{(A \cap \overline{B} \cap \overline{C}) \cup \mathrm{P}(\overline{\mathrm{A}} \cap B \cap \overline{C}) \cup P(\overline{A} \cap \overline{B} \cap C)\}$

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

$= (1/2)(2/3)(3/4) + (1/2)(1/3)(3/4) + (1/2)(2/3)(1/4)$

$= \frac{1}{4} + \frac{1}{8} + \frac{1}{12} = 11/24$.

5. A is known to hit the target in 2 out of 5 shots. B is known to hit the target in 3 out of 4 shots. Find the probability of the target being hit when both try.

Solution

$P(A) = \frac{2}{5}, P(B) = \frac{3}{4}$

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

As A and B are independent , $P(A \cap B) = P(A)P(B)$.

$\therefore P(A \cup B) = \frac{2}{5} + \frac{3}{4} - \left(\frac{2}{5}\right)\left(\frac{3}{4}\right)$

$= \frac{17}{20}$

6. Events A and B are such that $P(A + B) = 3/4, \mathrm{P(AB)} = 1/4, \mathrm{P}(\overline{\mathrm{A}}) = 2/3$ then find P(B).

Solution

$P(A) = 1 - P(\overline{A}) = 1 - 2/3 = 1/3$.

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

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

=(3/4)+(1/4)-(1/3)

-2/3.

7. If A and B are events with P(A) = 3/8, P(B)=1/2, $P(A \cap B) = 1/4$. Find $P(A^c \cap B^c)$

Solution

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

$= \frac{3}{8} + \frac{1}{2} - \frac{1}{4} = \frac{5}{8}$

$P(A^c \cap B^c) = P\left((A \cup B)^c\right) = 1 - P(A \cup B)$

=1-5/8

=3/8.

8. A lot consists of 10 good articles, 4 with minor defects and 2 with major defects. Two articles are chosen from the lot at random(without replacement). Find the probability that

a) both are good

b) both have major defects

c) at least 1 is good

d) at most 1 is good

e) exactly 1 is good

f) neither has major defects

g) neither is good

Although the articles may be drawn one after the other, we can consider that both articles are drawn simultaneously, as they are drawn without replacement.

Solution

a) $P(\text{both are good}) = \frac{\text{No. of ways drawing 2 good articles}}{\text{Total number of ways of drawing 2 articles}}$

$= \frac{10_{c_2}}{16_{c_2}} = \frac{3}{8}$.

b) $P(\text{both have major defects}) = \frac{\text{No. of ways of drawing 2 articles with major defects}}{\text{Total number of ways of drawing 2 articles}}$

$$= \frac{2_{c_2}}{16_{c_2}} = \frac{1}{120}$$

c) $P(\text{atleast 1 is good}) = P(\text{exactly 1 is good or both are good})$

$$= P(\text{exactly 1 is good and 1 is bad or both are good})$$

$$= \frac{10_{c_1} \times 6_{c_1} + 10_{c_2}}{16_{c_2}} = \frac{7}{8}$$

d) $P(\text{atmost 1 is good}) = \text{P(none is good or 1 is good and 1 is bad)}$

$$= \frac{10_{c_0} \times 6_{c_2} + 10_{c_1} \times 6_{c_1}}{16_{c_2}} = \frac{5}{8}$$

e) $P(\text{exactly 1 is good}) = \text{P(1 is good and 1 is bad)}$

$$= \frac{10_{c_1} \times 6_{c_1}}{16_{c_2}} = \frac{1}{2}.$$

f) $P(\text{neither has major defects}) = \text{P(both are non - major defective articles)}$

$$= \frac{14_{c_2}}{16_{c_2}} = \frac{91}{120}$$

g) $P(\text{neither is good}) = \text{P(both are defective)}$

$$= \frac{6_{c_2}}{16_{c_2}} = \frac{1}{8}$$

9. A box contains 4 white, 5 red and 6 black balls. Two balls are drawn at random. What is the probability that both are black?

Solution

Let A be the event of drawing two black balls.

Out of 15 balls, 2 balls can be selected in $^{15}c_2$ ways.

$\therefore n(S) = 15_{c_2} = 105$

Out of 6 balls, 2 balls can be selected in $^{6}c_2$ ways.

$\therefore n(A) = 6_{c_2} = 15$

$$P(A) = \frac{n(A)}{n(S)} = \frac{15}{105} = 0.1429$$

Hence the probability of drawing two black balls is 0.1429

10. From a well shuffled deck of 52 cards , 4 cards are selected at random. Find the probability that the selected cards are

i) 3 spades and 1 heart

ii) 2 kings, 1 ace and 1 queen

iii) all are diamonds

iv) There is one card of each suit

v) all the four are hearts and one of them is a jack

Solution

i) $$P(\text{getting 3 spades and 1 heart}) = \frac{13_{c_3} \times 13_{c_1}}{52_{c_4}} = 0.0137$$

ii) $$P(\text{getting 2 kings, 1 ace and 1 queen}) = \frac{4_{c_2} \times 4_{c_1} \times 4_{c_1}}{52_{c_4}} = 0.0003$$

iii) $$P(\text{getting all diamonds}) = \frac{13_{c_4}}{52_{c_4}} = 0.0026$$

iv) $$P(\text{getting one card from each suit}) = \frac{13_{c_1} \times 13_{c_1} \times 13_{c_1} \times 13_{c_1}}{52_{c_4}} = 0.1055$$

v) $$P(\text{getting 4 hearts out of which one is a jack}) = \frac{1_{c_1} \times 12_{c_3}}{52_{c_4}} = 0.0008$$

Exercises

1) Find the chance of throwing

i) four

ii) an even number with an ordinary six faced dice.

2) Find the probability of drawing an ace or a spade or both from a deck of cards

3) What is the probability of obtaining 2 heads in two throws of a single coin.

Conditional probability

If the probability of the event A provided the event B has already occurred is called the conditional probability and is defined as

$$P(A/B) = \frac{P(A \cap B)}{P(B)} \quad \text{provided} \quad P(B) \neq 0$$

If the probability of the event B provided the event A has already occurred is given by

$$\mathrm{P(B/A)} = \frac{\mathrm{P(A \cap B)}}{\mathrm{P(A)}} \quad \text{provided} \quad \mathrm{P(A)} \neq 0$$

Note

i) if the events A and B are independent then

$$\mathrm{P(A/B)} = \frac{\mathrm{P(A).P(B)}}{\mathrm{P(B)}} = \mathrm{P(A)}$$

$$\mathrm{P(B/A)} = \frac{\mathrm{P(A).P(B)}}{\mathrm{P(A)}} = P(B)$$

ii) If A and B are mutually exclusive events then

$$\mathrm{P(B/A)} = 0 \text{ and } \mathrm{P(A/B)} = 0 \text{ since } \mathrm{P(A \cap B)} = 0$$

Multiplication law of probability

If A and B are dependent events then

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

EXAMPLES

Total Probability Theorem

If $B_1, B_2, \ldots, B_n$ are mutually exclusive and exhaustive set of events of a sample space S and A is any event associated with the events $B_1, B_2, \ldots, B_n$ then

$$P(A) = P(B_1)P(A/B_1) + P(B_2)P(A/B_2) + \ldots + P(B_n)P(A/B_n) .$$

i.e., $$P(A) = \sum_{i=1}^{n} P(B_i)P(A/B_i) .$$

1) If the probability that a communication system will have high fidelity 0.81 and the probability that it will have high fidelity and selectivity is 0.18, what is the probability that a system with high fidelity will also have selectivity?

Solution

Let A be the event that the system has selectivity.

B be the event that the system has fidelity

$P(B) = 0.81, \mathrm{P(A \cap B)} = 0.18$

$$P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{0.18}{\text{o.81}} = \frac{2}{9}$$

2)A box contains 4 bad and 6 good tubes. Two are drawn out from the box at a time. One of them tested and found to be good. What is the probability that the other one also is good?

Solution

Let A = one of the tubes drawn is good and B= the other tube is good

P(A)=6/10

$$P(A \cap B) = P(\text{both tubes drawn are good}) = \frac{6_{c_2}}{10_{c_2}} = \frac{1}{3}$$

Knowing that one tube is good, the conditional probability that the other tube is also good is required, i.e., P(B/A) is required.

By definition,

$$P(B/A) = \frac{P(A \cap B)}{P(A)} = \frac{1/3}{6/10} = \frac{5}{9}$$

3) Two defective tubes get mixed up with 2 good ones. The tubes are tested, one by one until both defectives are found. What is the probability that the last defective tube is obtained on

a) the second test

b) the third test and

c) the fourth test

Solution

Let D represent defective and N represent non-defective tube.

a) $P(\text{Second D in the II test}) = \mathrm{P(D \text{ in the I test and D in the II test})}$

$= P(D_1 \cap D_2), \text{say}$

$= P(D_1)P(D_2)$

$$= \frac{2}{4} \times \frac{1}{3} = \frac{1}{6}$$

b) $P(\text{Second D in the III test}) = \text{P}(\text{D}_1 \cap \text{N}_2 \cap \text{D}_3 \text{ or } \text{N}_1 \cap \text{D}_2 \cap \text{D}_3)$

$$= \text{P}(\text{D}_1 \cap \text{N}_2 \cap \text{D}_3) + P(\text{N}_1 \cap \text{D}_2 \cap \text{D}_3)$$

$$= \frac{2}{4} \times \frac{2}{3} \times \frac{1}{3} + \frac{2}{4} \times \frac{2}{3} \times \frac{1}{2} = \frac{1}{3}$$

$P(\text{Second D in the IV Test}) = \text{P}(\text{D}_1 \cap \text{N}_2 \cap N_3 \cap \text{D}_4 \text{ or } \text{N}_1 \cap \text{D}_2 \cap N_3 \cap \text{D}_4 \text{ or } \text{N}_1 \cap \text{N}_2 \cap D_3 \cap \text{D}_4)$
$= \text{P}(\text{D}_1 \cap \text{N}_2 \cap N_3 \cap \text{D}_4) + \text{P}(\text{N}_1 \cap \text{D}_2 \cap N_3 \cap \text{D}_4) + \text{P}(\text{N}_1 \cap \text{N}_2 \cap D_3 \cap \text{D}_4)$

$$= \frac{2}{4} \times \frac{2}{3} \times \frac{1}{2} \times 1 + \frac{2}{4} \times \frac{2}{3} \times \frac{1}{2} \times 1 + \frac{2}{4} \times \frac{2}{3} \times \frac{1}{2} \times 1$$
$$= \frac{1}{2}$$

4) In a shooting test, the probability of hitting the target is 1/2 for A, 2/3 for B and 3/4 for C. If all of them fire at the target, find the probability that

a) none of them hits the target

b) atleast one of them hits the target.

Solution

$$P(\overline{A}) = 1 - P(A) = 1 - 1/2 = 1/2$$

$$P(\overline{A}) = \frac{1}{2} \text{P}(\overline{\text{B}}) = \frac{1}{3}, P(\overline{C}) = \frac{1}{4}$$

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

$$= \frac{1}{2} \times \frac{1}{3} \times \frac{1}{4} = \frac{1}{24}$$

P(at least one hits the target)$= 1 - P(\text{none hits the target})$

$$= 1 - \frac{1}{24} = \frac{23}{24}$$

5) A box contains tags marked 1,2,…,n two tags are chosen at random without replacement. Find the probability that the numbers on the tags will be consecutive integers.

Solution

Number of ways of choosing any one pair from the (n-1) pairs = $(n-1)_{c_1} = n-1$

Total number of ways of choosing 2 tags from the n tags = n_{c_2}

$\therefore$ The required probability = $\dfrac{n-1}{n_{c_2}}$

6) Among the workers in a factory only 30% receive a bonus. Among those receiving the bonus only 20% are skilled. What is the probability of a randomly selected worker who is skilled and receiving bonus?

Solution

P(A)=30/100 =0.3

$$P(B/A) = \frac{20}{100} = 0.2$$

$$P(A \cap B) = P(A)P(B/A) = (0.3)(0.2) = 0.06$$

7) A and B alternately throw a pair of dice. A wins if he throws 6 before B throws7 and B wins if he throws 7 before A throws 6. If A begins, show that his chance of winning is 30/61.

Solution

Let A be the event of A throwing 6 and B be the event of B throwing 7.

$$P(A) = \frac{5}{36}, \mathrm{P(B)} = \frac{1}{6}$$

$$P(A \text{ wins}) = P(\text{A or ABA or ABABA or...})$$
$$= \mathrm{P(A)} + \mathrm{P(ABA)} + \mathrm{P(ABABA)} + ...$$

$$= \frac{5}{36} + (\frac{5}{36})(\frac{1}{6})(\frac{5}{36}) + ... = \frac{30}{61}$$

8) In a coin tossing experiment, if the coin shows head, 1 dice is thrown and the result is recorded. But if the coin shows tail, 2 dice are thrown and their sum is recorded. What is the probability that the recorded number will be 2?

Solution

When a single dice is thrown,

$$P(\text{getting } 2) = 1/6$$

When 2 dice are thrown, the sum will be 2, only if each dice shows 1.

$\therefore P(\text{getting 2 as sum with 2 dice}) = \text{P(getting 1 in first dice)p(getting 1 in second dice)}$

$$= \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$$

By theorem of total probability,

$$P(A) = \sum_{i=1}^{n} P(B_i) P(A / B_i)$$

P(recorded number will be 2) = P(H)P(getting 2 with 1 dice)+P(T)P(getting 2 as a sum with 2 dice)

$$= (1/2)(1/6) + (1/2)(1/36)$$

$$= \frac{6+1}{72} = \frac{7}{72}$$

9) If atleast 1 child in a family with 2 children is a boy, what is the probability that both children are boys?

Solution

S= {BB.BG.GB,BB}

P(atleast 1 child is a boy)=P(exactly 1 boy)+P(2 boys)

$$= \frac{2}{4} + \frac{1}{4} = \frac{3}{4}$$

Therefore the probability that both children are boys given that atleast 1 child is a boy=

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

$$= \frac{1/4}{3/4} = 1/3$$

10) Two fair dice are thrown independently. Three events A, B and C are defined as follows.

i) odd face with the first dice

ii) odd face with the second dice

iii) sum of the numbers in 2 dice is odd.

Are the events A, B and C mutually independent?

Solution

$P(A) = \frac{3}{6} = 1/2, \text{P(B)} = 1/2, \text{P(C)} = 1/2$

$P(A \cap B) = P(A)P(B) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$

$P(A \cap B) = P(B \cap C) = P(A \cap C) = 1/4$

$P(A \cap B \cap C) = 0$, since C can not happen when A and B occur.

Therefore $P(A \cap B \cap C) \neq P(A)P(B)P(C)$

Therefore the events are pairwise independent, but not mutually independent.

11) A bold is manufacture by 3 machines A, B and C. A turns out twice as many items as B, and machines B and C produce equal number of times. 2% of bolts produced by A and B are defective and 4% of bolts produced by C are defective. All bolts are put into 1 stock pile and 1 is chosen from this pile. what is the probability that it is defective?

Solution

Let A, B and C be the event in which the item has been produced by machine A, B and C.

Let D be the event of the item to be defective.

P(A)=1/2, P(B)= P(C)=1/4

P(D/A)=P(an item is defective, given that A has produced)=2/100

P(D/B)=P(an item is defective, given that B has produced)=2/100

P(D/C)=P(an item is defective , given that C has produced) =4/100

By theorem on total probability,

$P(D) = P(A)P(D/A) + P(B)P(D/B) + P(C)P(D/C)$

$P(D) = (1/2)(2/100) + (1/4)(2/100) + (1/4)(4/100)$
$= 1/40$

12) A bag contain 5 red and 3 green balls and a second bag contains 4 red and 5 green balls. One of the bags is selected at random and a draw of 2 balls is made from it. What is the probability that one of them red and the other is green.

Solution

Let A1 and and A2 denote the first bag and second bag respectively. Then

$P(A_1)=P(A_2)=1/2$

Let B be the event of selecting one red and one green ball.

$P(B/A_1) = \frac{5_{c_1} \times 3_{c_1}}{8_{c_2}}$ 15/28

P(B/A2)= $\frac{4_{c_1} \times 5_{c_1}}{9_{c_2}} = \frac{20}{36} = \frac{5}{9}$

$\therefore$ The required probability = $P(A_1)P(B/A_1)+P(A_2)P(B/A_2)$

$$= (1/2)(15/28) + (1/2)(5/9) = \frac{15}{56} + \frac{5}{18}$$

=275/504

13) An urn contains 10 white and 3 black balls. Another urn contains 3 white and 5 black balls. Two balls are drawn at random from the first urn and placed in the second urn and then 1 ball is taken at random from the latter. What is the probability that it is a white ball?

Solution

The two balls transferred may be both white or both black or 1 white and 1 black.

Let B_1 be the event of drawing 2 white balls from the first urn and B_2 be the event of drawing 2 black balls from it and B_3 be the event of drawing 1 white and 1 black ball from it.

Let A be the event of drawing a white ball from the second urn after transfer.

$$P(B_1) = 15/26, \mathrm{P}(\mathrm{B}_2) = 1/26, \mathrm{P}(\mathrm{B}_3) = 10/26$$

$$P(A/B_1) = P(\text{drawing a white ball / 2 white balls have been transferred})$$
$$= 5/110$$

Similarly $P(A/B_2) = 3/10$ and $P(A/B_3) = 4/10$

Therefore, $P(A) = P(B_1) \times P(A/B_1) + P(B_2)P(A/B_2) + P(B_3)P(A/B_3)$

$$= \left(\frac{15}{26} \times \frac{5}{10}\right) + \left(\frac{1}{26} \times \frac{3}{10}\right) + \left(\frac{10}{26} \times \frac{4}{10}\right) = \frac{75+3+40}{260} = \frac{118}{260}$$

=59/130

Baye's Theorem

If $B_1, B_2, \ldots, B_n$ are a set of exhaustive and mutually exclusive events of a sample space S and A is any event associated with $B_1, B_2, \ldots, B_n$ such that

$$A \subseteq \bigcup_{i=1}^{n} B_i \qquad \text{then} \qquad P(B_i/A) = \frac{P(B_i)P(A/B_i)}{\sum_{i=1}^{n} P(B_i)P(A/B_i)}$$

Examples:

1). The contents of Urn I, II and III are as follows:

i) 1 white, 2 black and 3 red balls

ii) 2 white, 1 black and 1 red balls and

iii) 4 white, 5 black and 3 red balls

One urn is chosen at random and two balls are drawn. They happen to be white and red. What is the probability that they come from urn I, urn II and III?

Solution

There are 3 urns. The probability of choosing one urn is 1/3.

Let B_1 be the event of choosing urn I, B_2 be the event of choosing urn II and B_3 be the event of choosing urn III.

$P(B_1) = 1/3, \mathrm{P}(\mathrm{B}_2) = 1/3, P(B_3) = 1/3$.

Let A be the event of choosing 2 balls are white and one red. If the urn I is chosen then

$$P(A/B_1) = \frac{1_{c_1} \times 3_{c_1}}{6_{c_2}} = 1/5$$

.

If the urn II is chosen, then

$$P(A/B_2) = \frac{2_{c_1} \times 1_{c_1}}{4_{c_2}} = \frac{2}{6} = 1/3$$

If the urn III is chosen

$$P(A/B_3) = \frac{4_{c_1} \times 3_{c_1}}{12_{c_2}} = 2/11$$

.

$$\therefore P(A) = P(B_1)P(A/B_1) + P(B_2)P(A/B_2) + P(B_3)P(A/B_3)$$

$$= \frac{1}{3} \times \frac{1}{5} + \frac{1}{3} \times \frac{1}{3} + \frac{1}{3} \times \frac{2}{11} = \frac{118}{495}$$

Assume that A has happened

i.e., 1 white and 1 red balls are chosen.

The probability that they come from urn I is

$$P(B_1/A) = \frac{P(B_1)P(A/B_1)}{P(A)}$$

$$= \frac{\frac{1}{3} \times \frac{1}{5}}{\frac{118}{495}}$$

=0.2797.

The probability that they come from urn II is

$$P(B_2 / A) = \frac{P(B_2)P(A/B_2)}{P(A)} = \frac{\frac{1}{3}\times\frac{1}{3}}{118/495} - 0.4661$$

.

The probability that they come from urn III is

$$P(B_3 / A) = \frac{P(B_3)P(A/B_3)}{P(A)} = \frac{\frac{1}{3}\times\frac{2}{11}}{118/495} - 0.2542$$

.

2) A bag contains 5 balls and it is not known how many of them are white. Two balls are drawn at random from the bag and they are noted to be white. What is the chance that all the balls in the bag are white?

Solution

Since 2 white balls have been drawn out, the bag must have contained 2,3,4 or 5 white balls.

Let B_1 be the event of the bag containing 2 white balls, B_2 be the event of the bag containing 3 white balls, B_3 be the event of the bag containing 4 white balls and B_4 be the event of the bag containing 5 white balls.

Let A be the event of drawing 2 white balls.

$$P(A/B_1) = 1/10, \mathrm{P(A/B_2)} = 3/10, \mathrm{P(A/B_3)} = 3/5, \mathrm{P(A/B_4)} = 1$$

$P(B_1)=P(B_2)=P(B_3)=P(B_4)=1/4$

By Bayes theorem

$$P(B_i / A) = \frac{P(B_i)P(A/B_i)}{\sum_{i=1}^{n} P(B_i)P(A/B_i)}$$

P(all the balls in the bag are white)=$P(B_4/A)$

$$P(B_4 / A) = \frac{P(B_4)P(A/B_4)}{\sum_{i=1}^{4} P(B_i)P(A/B_i)}$$

$$= \frac{(1/4)(1)}{(1/4)(1/10)+(1/4)(3/10)+(1/4)(3/5)+(1/4)1}$$

$$= \frac{1/40}{(\frac{1}{40}+\frac{3}{40}+\frac{3}{20}+\frac{1}{4})} = \frac{1/40}{20/40} = 1/20$$

3) In a bolt factory, machines A, B and C produce 25%, 35% and 40% of the total output respectively. Of their outputs 5,4 and 2% respectively are defective bolts. If a bolt is chosen at random from the combined output, what is the probability that it is defective? If a bolt chosen at random is found to be defective, what is the probability that it was produced by B?

Solution

$P(E_1)=0.25$, $P(E_2)= 0.35$, $P(E_3)=0.40$

Let X be the event of drawing defective dolt.

$P(X/E_1)=0.05$

$P(X/E_2)=0.04$

$P(X/E_3)= 0.02$

By Bayes theorem

$$P(E_2 / X) = \frac{P(E_2)P(X / E_1)}{P(E_1)P(X / E_{1)} + P(E_2)P(X / E_2) + P(E_3)P(X / E_3)}$$

$$= 0.406$$

4) A given lot of IC chips contains 2% defective chips. Each is tested before delivery. The tester itself is not totally reliable. Probability of tester says the chip is good when it is really good is 0.95 and the probability of tester says chip is defective when it is actually defective is 0.94. If a tested device is indicated to be defective, what is the probability that it is actually defective?

Solution

Let E be the event of chip which is actually good and D be the event of tester says it is good.

$$P(\overline{E}) = 0.02$$

$$P(E) = 1 - P(\overline{E}) = 1 - 0.02 = 0.98$$

Given that the probability of tester says the chip is good when it is really good is 0.95.

$$P(D / E) = 0.95$$

$$P(\overline{D} / E) = 1 - 0.95 = 0.05$$

$$P(D / \overline{E}) = 0.94$$

$$P\left(\overline{D}/\overline{E}\right)=1-0.94=0.06$$

$$P(\overline{E}/\overline{D})=\frac{P(\overline{E})P(\overline{D}/\overline{E})}{P(E)P(\overline{D}/E)+P(\overline{E})P(\overline{D}/\overline{E})}$$

$$=\frac{(0.02)(0.94)}{(0.98)(0.05)+(0.02)(0.94)}$$

$$=\frac{0.0188}{0.0678}$$

$$=0.2773$$

5) A certain firm has plant A, B and C producing IC chips. Plant A produces twice the output from B and B produces twice the output from C. The probability of a non defective product produced by A,B and C are respectively 0.85, 0.75, and 0.95. A customer receives a defective product, find the probability that it came from plant B.

Solution:

P(A)=1; P(B)=0.5; P(C)=0.25

$P(E/A)=0.85$; P(E/B) = 0.75; P(E/C) = 0.95

$P(\overline{E}/A)=0.15$; $P(\overline{E}/B)=0.25$; $P(\overline{E}/C)=0.05$

The probability that the customer receives a defective product from plant B is

is

$$P(B/\overline{E})=\frac{P(B)P(\overline{E}/B)}{P(A)P(\overline{E}/A)+P(B)P(\overline{E}/B)+P(C)P(\overline{E}/C)}$$

$$=\frac{(0.5)(0.25)}{1(0.15)+(0.5)(0.25)+(0.25)(0,05)}$$

$$=\frac{0.125}{0.2875}=0.4348$$

6) There are 3 true coins and 1 false coin with head on both sides. A coin is chosen at random and tossed 4 times. If head occurs all the 4 times, what is the probability that the false coin has been chosen and used?

Solution

P(T)=P(the coin is a true coin)=3/4

P(F)=P(the coin is a false coin)=1/4

Let A be the event of getting all heads in 4 tosses.

P(A/T)= $\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}=\frac{1}{16}$

P(A/F)=1

By Bayes theorem

$$P(F/A)=\frac{P(F)\times P(A/F)}{P(F)P(A/F)+P(T)P(A/T)}$$

$$= \frac{\frac{1}{4} \times 1}{\frac{1}{4}(1) + (\frac{3}{4})(\frac{1}{16})}$$

$$= \frac{16}{19}$$

Random Variable:

A random variable X whose value is determined by the outcome of a random experiment is called a random variable.

Example

A random experiment consists of two tosses of a coin. Consider the random variable which is the number of heads (0,1 or 2)

Outcome	H H	H T	T H	T T
Value of X	2	1	1	0

If the function values are ordered pairs of real numbers, the function is said to be a two-dimensional real numbers

Discrete Random Variable

A random variable which can assume only a countable number of real values is called a discrete random variable

Example

Number of telephone calls per unit time, marks obtained in a test, number of printing mistakes in each page of a book.

Probability Mass Function

Suppose X is an one-dimensional discrete random variable taking atmost a countably infinite number of values $x_1, x_2, \ldots$ with each possible outcome x_i , we associate a number p_i, $P(X=x_i)=p(x_i)=p_i$, called the probability of x_i.

The function $p(x_i)$,i-1,2… satisfying the conditions

i. $p(x_i) \geq 0 \forall i$

ii. $\sum_{i=1}^{\infty} p(x_i) = 1$

is called the probability mass function or probability function of the random variable X.

The collection of pairs $\{x_i, p_i\}$, I =1,2,3… is called the probability distribution of the random variable X.

Discrete Distribution Function

The distribution function of the random variable X with probability mass function $p(x_i)$, i=1,2,…,n is defined as $F(x_i) = \sum_{i: x_i \le x} p(x_i)$.

Note:

i) $p(x_i) = \sum_{i: x_i \le x} P(X = x_i)) = F(x_i) - F(x_{i-1})$

where F is the distribution of the random variable X.

ii) Mean of the random variable X = $E(X) = \sum_x xp(x)$

iii) Variance of the random variable X = $V(X) = \sum x^2 p(x) - \left[\sum xp(x)\right]^2$

$$V(X) = E(X^2) - [E(X)]^2$$

Continuous Random Variable

A random variable X is said to be continuous if it can take all possible values between certain limits.

Probability density function

The probability density function $f_X(x)$ of a continuous random variable X is defined as $P(x \le X \le x + dx) = f_X(x)dx$ where (x,x+dx) is an infinitesimally small interval and satisfies the following conditions.

i. $f_X(x)$ is integerable over the range $-\infty \le x \le \infty$
ii. $f(x) \ge 0$ for all x, $-\infty \le x \le \infty$
iii. $\int_{-\infty}^{\infty} f(x)dx = 1$.

Properties of Probability density function

The probability density function $f_X(x)$ of a continuous random variable X satisfies the following properties

i. $f(x) \geq 0$

ii. $\int_{-\infty}^{\infty} f(x)dx = 1$

iii. $P(x_1 < X \leq x_2) = \int_{x_1}^{x_2} f_X(x)dx$

iv. $P(x_1 < X \leq x_2) = P(x_1 < X < x_2)$

v. $P(X = a) = \int_{-a}^{a} f(x)dx = 0$

[Note: In case of a continuous random variable, the probability at a point is always zero.

i.e, P(X=a)=0 for all possible values of a.

Cumulative Distribution Function

The cumulative distribution F(x) of a continuous random variable X with PDF f(x) is given by

$$F(x) = P(X \leq x) = \int_{-\infty}^{\infty} f(x)dx, \ -\infty < x < \infty$$

Note:

i. $P(a < X < b) = F(b) - F(a)$

ii. $F(-\infty) = \underset{x \to -\infty}{Lt} F(x) = 0 \ ; \ F(\infty) = \underset{x \to \infty}{Lt} F(x) = 1$

iii. The relation between the CDF and PDF is

iv. $f(x) = \frac{d}{dx} F(x)$

v. If X is a continuous random variable with PDF f(x) then

$$\text{Mean} = \text{E(X)} = \int_{-\infty}^{\infty} \text{xf(x)dx}$$

Mean =

$$\text{E}(\text{X}^2) = \int_{-\infty}^{\infty} \text{x}^2\text{f(x)dx}$$

$$\text{Var(X)} = \text{E}(\text{X}^2) - [E(X)]^2$$

Examples:

1) A random variable X has the following probability distribution

X	0	1	2	3	4	5	6	7	8

P(X)	A	3a	5a	7a	9a	11a	13a	15a	17a

Find

i) a

ii) $P(X \leq 2)$

iii) The distribution function of X and

iv) the mean of X

Solution

i) If p(x) is the probability mass function, then $\sum p(x) = 1$

i.e., a+3a+5a+7a+9a+11a+13a+15a+17a=1

81a=1

a=1/81

ii) $P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2)$ a=99

=a+3a+5a

=9a

=9(1/81)=1/9.

iii) The distribution function of X is

$F(x) = 0, x < 0$

$= P(X = 0) = a = \frac{1}{81}, 0 \le \text{x} < 1$

$= \text{P(X} \le 1) = 4\text{a} = \frac{4}{81}, 1 \le \text{x} < 2$

$= \text{P(X} \le 2) = 9\text{a} = \frac{9}{81}, 2 \le x < 3$

$= P(X \le 3) = 16a = 16(\frac{1}{81}), 3 \le x < 4$

$= P(X \le 4) = 25a = 25(\frac{1}{81}), 4 \le x < 5$

$= P(X \le 5) = 36a = 36(\frac{1}{81}), 5 \le x < 6$

$= P(X \le 6) = 49a = 49(\frac{1}{81}), 6 \le x < 7$

$= P(X \le 7) = 64a = 64(\frac{1}{81}), 7 \le x < 8$

$= P(X \le 8) = 81a = 81(\frac{1}{81}) = 1, 8 \le x \; i.e., x \ge 8$

iv) The mean of X

i. e., $E(X) = \sum xp(x)$

=0+3a+10a+21a+a+36a+55a+78a+105a+136a

=444a

=444(1/81)

=5.48

$\text{Var(X)} = \sum \text{x}^2 p(x) - \left[\sum xp(x)\right]^2$

2. If the random variable X takes the values 1,2,3 and 4 such that $2P(X = 1) = 3P(X = 2) = P(X = 3) = 5P(X = 4)$, find the probability distribution function and cumulative distribution function of X.

Solution

Let $P(X = 3) = k$ then

$2P(X = 1) = k \Rightarrow \text{P(X} = 1) = \frac{\text{k}}{2}$

$3P(X = 2) = k \Rightarrow \text{P(X} = 2) = \frac{\text{k}}{3}$

$$5P(X=4)=k \Rightarrow \mathrm{P(X=4)}=\frac{\mathrm{k}}{5}$$

$$\sum_{i=1}^{4} p(x_i)=1 \Rightarrow \frac{k}{2}+\frac{k}{3}+k+\frac{k}{5}=1$$

$$k=\frac{30}{61}$$

The probability distribution function

$$\mathrm{P(X=1)}=\frac{\mathrm{k}}{2}=\frac{30/61}{2}=0.2459$$

$$\mathrm{P(X=2)}=\frac{\mathrm{k}}{3}=\frac{30/61}{3}=\frac{30}{183}=0.1639$$

$$\mathrm{P(X=3)}=k=\frac{30}{61}=0.491$$

$$\mathrm{P}(X=4)=\frac{\mathrm{k}}{5}=\frac{30/61}{5}=0.0984$$

To find Cumulative distribution function

The CDF F(x)= $P(X \le x)$ is calculated as follows

$F(x)=0, \mathrm{x}<1$
$=0.2459, 1 \le \mathrm{x}<2$
$=0.2459+0.1639=0.4098, 2 \le x<3$
$=0.4098+0.491=0.9008, 3 \le x<4$
$=0.996=1, x \ge 4$

3). The probability distribution function of X is

X=i	1	2	3	4
P(X=i)	15/61	10/61	30/61	6/61

Find cumulative distribution function of X

Solution

The CDF F(x)= $P(X \le x)$ is defined as follows

$F(x) = 0, \text{x} < 1$

$= \frac{15}{61}, 1 \le \text{x} < 2$

$= \frac{25}{61}, 2 \le x < 3$

$= \frac{55}{61}, 3 \le x < 4$

$= 1, x \ge 4$

4) A continuous random variable has the probability density function

$f(x) = k(x-1)^3, 1 \le \text{x} \le 3$

$= 0, \text{otherwise}$

Find

i) the value of k

ii) the distribution function F(x)

Solution

Since $\int_{-\infty}^{\infty} f(x)dx = 1$ we have

$$\int_1^3 k(x-1)^3 dx = 1$$

$$= k\frac{(x-1)^4}{4}\Big]_1^3 = 1$$

$$(k/4)(2^4 - 0) = 1$$

$$k = 1/4$$

$$\therefore f(x) = 1/4(x-1)^3, 1 \le x \le 3$$

ii) $F(x) = \int_{-\infty}^{x} f(x)dx$

$$= \int_{-\infty}^{1} f(x)dx + \int_1^x f(x)dx$$

$$= \int_{-\infty}^{1} 0.dx + \int_{1}^{3} \frac{1}{4}(x-1)^3 dx$$

$$= \frac{1}{16}(x-1)^4$$

$$\therefore F(x) = 0, \text{x} < 1$$

$$= \frac{1}{6}(x-1)^4, 1 \le \text{x} \le 3$$

$$= 1, \text{x} > 3$$

5) A random variable X has the PDF f(x) given by

$$f(x) = cxe^{-x}, \text{x} > 0$$
$$= 0, \text{x} \le 0$$

Find the value of c and CDF OF X

Solution

If f(x) is a PDF, then

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

$$\int_{0}^{\infty} cxe^{-x}dx = 1$$

$$c\left[-xe^{-x} + (-e^{-x})\right]_0^{\infty} = 1$$

$$\Rightarrow c(1) = 1$$

c=1

$$\therefore f(x) = xe^{-x}$$

$$P(X \le x) = \int_{0}^{\infty} xe^{-x}dx$$

The CDF of X=F(x)=

$$\Rightarrow F(x) = -xe^{-x} - e^{-x}\Big]_0^x$$

$$= xe^{-x} - e^{-x} + 1$$

$= 1 - e^{-x}(1+x)$

$\therefore F(x) = 1 - e^{-x}(1+x)$, x > 0
$= 0$, otherwise

6) A continuous random variable X follows the probability law f(x)=ax², $0 \le x \le 1$. Determine a and find the probability that x lies between $\frac{1}{4}$ and $\frac{1}{2}$.

Solution

If f(x) is a PDF then

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

$$\int_{-\infty}^{0} f(x)dx + \int_{0}^{1} f(x)dx + \int_{1}^{\infty} f(x)dx = 1$$

$$\int_{0}^{1} ax^2 dx = 1$$

$$\Rightarrow a\frac{x^3}{3}\Big]_0^1 = 1$$

$$\Rightarrow a\left[\frac{1}{3} - 0\right] = 1$$ i.e., a=3

$\therefore f(x) = 3x^2$

$$P\left[\frac{1}{4} < x < \frac{1}{2}\right] = \int_{1/4}^{1/2} f(x)dx$$

$$= \int_{1/4}^{1/2} 3x^2 dx$$

$$= 3\frac{x^3}{3}\Big]_{1/4}^{1/2} = \left(\frac{1}{2}\right)^3 - \left(\frac{1}{4}\right)^3$$

$$= \frac{1}{8} - \frac{1}{64} = \frac{7}{64}$$

7) If the PDF of a random variable x is f(x)=x/2 in $0 \le x \le 2$. Find $P(X > 1.5 / X > 1)$

Solution

$$P(X > 1.5 / X > 1) = \frac{P(X > 1.5)P(X > 1.5 \cap X > 1)}{P(X > 1)}$$

$$= \frac{\int_{1.5}^{2} f(x)dx}{\int_{1}^{2} f(x)dx}$$

$$= \frac{\int_{1.5}^{2} \frac{x}{2}dx}{\int_{1}^{2} \frac{x}{2}dx}$$

$$= \frac{\frac{1}{2}\left(\frac{x^2}{2}\right)\Big]_{1.5}^{2}}{\frac{1}{2}\left(\frac{x^2}{2}\right)\Big]_{1.}^{2}}$$

$$= \frac{4-(1.5)^2}{(4-1)}$$

$$= \frac{4-2.25}{3}$$

$$= \frac{1.75}{3} = 0.5833$$

8) If X is a continuous random variable with PDF

$f(x) = x, 0 \le \text{x} < 1$

$= \frac{1}{2}(x-1)^2, 1 \le x \le 2$

$= 0, \text{otherwise}$ (x)

Find the cumulative distribution function F(x) of X and use it to find $P(\frac{3}{2} < x < \frac{5}{2})$

Solution

By definition $F(x) = P(X \le x)$

$$F(x) = \int_0^1 x dx + \int_1^x \frac{3}{2}(x-1)^2 dx$$

$$= \frac{x^2}{2}\Big]_0^1 + \frac{3}{2}\frac{(x-1)^3}{3}\Big]_1^x$$

$$= \frac{1}{2} - 0 - \frac{(x-1)^3}{2}$$

$$= \frac{1}{2} + \frac{x^3 - 3x^2 + 3x - 1}{2}$$

$$\frac{x^3 - 3x^2 + 3x}{2}, 1 \le x \le 2$$

$F(x) = 1$ when $x > 2$

The cumulative distribution function is

$$F(x) = x^2/2, 0 \le x \le 1$$
$$= \frac{x^3 - 3x^2 + 3x}{2}, 1 \le x \le 2$$
$$= 1 \text{ when } x > 2$$

To find $P(\frac{3}{2} < x < \frac{5}{2})$.

$$P(\frac{3}{2} < x < \frac{5}{2}) = F(5/2) - F(3/2)$$
$$= 1 - \frac{(3/2)^3 - 3(3/2)^2 + 3(3/2)}{2}$$

$$= 1 - \frac{(27/8) - (27/4) + (9/2)}{2}$$

$$= 1 - \frac{1}{2}\left(\frac{27 - 54 + 36}{8}\right)$$

=1-(9/16)=7/16.

9) If the density function of a continuous random variable X is

$f(x) = ax, 0 \le x \le 1$

$= a, 1 \le x \le 2$

$= 3a - ax, 2 \le x \le 3$

$= 0, \text{otherwise}$

i) Find the value of a

ii) the cumulative distribution function of X and

iii) $P(X \le 1.5)$

Solution

Since f(x) is a PDF of x, we have $\int_{-\infty}^{\infty} f(x)dx = 1$

$$\int_0^1 axdx + \int_1^2 adx + \int_2^3 (3a - ax)dx + 0 = 1$$

$$\Rightarrow a\frac{x^2}{2}\Big)_0^1 + ax]_1^2 + \Big[3ax - a\frac{x^2}{2}\Big]_2^3 = 1$$

$$\Rightarrow \frac{a}{2}(1-0) + a(2-1) + (3a.3 - a\frac{3^2}{2} - 6a + a.\frac{4}{2}) = 1$$

$$\Rightarrow \frac{a}{2} + a + (9a - \frac{9a}{2} - 6a + 2a) = 1$$

$$\Rightarrow \frac{a}{2} + a + (a/2) = 1$$

$$\frac{4a}{2} = 1$$

a=1/2.

The cumulative distribution function of the random variable X is F(x)= $P(X \le x)$

$F(x) = 0, x < 0$

$= \int_0^x axdx, 0 \le x \le 1$

$= a\frac{x^2}{2}\Big]_0^x = \frac{a}{2}x^2, 0 \le x \le 1$

$$F(x)=\int_0^1 axdx+\int_1^x adx$$

$$a\frac{x^2}{2}\Big]_0^1+ax\Big]_1^x=\frac{a}{2}+ax-a=-\frac{a}{2}+ax=a(x-\frac{1}{2})$$

$$a\frac{x^2}{2}\Big]_0^1+ax\Big]_1^x=\frac{a}{2}+ax-a=-\frac{a}{2}+ax=a(x-\frac{1}{2})=\frac{1}{2}(x-\frac{1}{2})$$

$$=\frac{x}{2}-\frac{1}{4},1\le x\le 2$$

$$F(x)=\int_0^1 axdx+\int_1^2 adx+\int_2^x (3a-ax)dx$$

$$F(x)=a\frac{x^2}{2}\Big]_0^1+ax\Big]_1^2+(3ax-a\frac{x^2}{2})\Big]_2^x$$

$$=\frac{a}{2}(1-0)+(2a-a)+(3ax-a\frac{x^2}{2}-6a+2aa)$$

$$=\frac{a}{2}+a+3ax-a\frac{x^2}{2}-4a$$

$$=\frac{-5a}{2}+3ax-a\frac{x^2}{2}$$

$$=\frac{-5}{4}+3\frac{x}{2}-\frac{x^2}{4},2\le \mathrm{x}\le 3$$

F(x)=1 when $x>3$

$\therefore$ The CDF of f(x) is

$$F(x)=\frac{1}{4}x^2,0\le \mathrm{x}\le 1$$

$$=\frac{x}{2}-\frac{1}{4},1\le \mathrm{x}\le 2$$

$$=-\frac{5}{4}+\frac{3x}{2}-\frac{x^2}{4},2\le \mathrm{x}\le 3$$

$=1$ when $\mathrm{x}>3$

$$P(X \le 1.5) = F(1.5) = \frac{1.5}{2} - \frac{1}{4} = \frac{1}{2}$$

The CDF of a random variable X is given by

$$F(x) = 0, x < 0$$
$$= \mathrm{x}^2, 0 \le \mathrm{x} < \frac{1}{2}$$
$$= 1 - \frac{3}{25}\left(3 - x^2\right), \frac{1}{2} \le \mathrm{x} < 3$$
$$= 1 \text{ when } \mathrm{x} \ge 3$$

7) Find the PDF of X and evaluate $P(|X| \le 1)$ and $P(\frac{1}{3} \le X \le 4)$ using both PDFand CDF

Solution

Given the C.D.F. of F(x) is

$$F(x) = 0, x < 0$$
$$= \mathrm{x}^2, 0 \le \mathrm{x} < \frac{1}{2}$$
$$= 1 - \frac{3}{25}\left(3 - x^2\right), \frac{1}{2} \le \mathrm{x} < 3$$
$$= 1 \text{ when } \mathrm{x} \ge 3$$

W.K.T

$$f(x) = \frac{d}{dx}\left[F(x)\right] = F'(x)$$

The PDF of X is

$$f(x) = 0, x < 0$$
$$= 2x, 0 \le \mathrm{x} < 1/2$$
$$= 1 - \frac{3}{25} 2(3 - x)(-1), \frac{1}{2} \le x < 3$$
$$= 0, x \ge 3$$

Case i: Using PDF

$$p(|X| \le 1) = p(-1 \le X \le 1) = \int_{-1}^{1} f(x)dx = \frac{13}{25}$$

$$p(\frac{1}{3} \le X \le 4) = \int_{1/3}^{4} f(x)dx = \frac{8}{9}$$.

Case ii: Using the CDF

$$P(|X| \le 1) = P(-1 \le X \le 1) = F(1) - F(-1)$$

$$= 1 - \frac{3}{25}(2^2)$$

$$= 1 - \frac{12}{25}$$

$$p(|X| \le 1) = 13/25.$$

$$P(1/3 \le X \le 4) = F(4) - F(1/3)$$

$$= 1 - \frac{1}{3^2} = \frac{8}{9}$$

8) If the random variable X has the PDF

$$f(x) = 1/4, |x| < 2$$
$$= 0, \text{otherwise}$$

Find

i. $P(X < 1)$
ii. $P(|X| > 1)$
iii. $P(2X + 3 > 5)$

Solution

$$P(X < 1) = \int_{-\infty}^{1} f(x)dx$$

$$= \int_{-\infty}^{-2} \frac{1}{4}dx + \int_{-2}^{1} \frac{1}{4}dx \quad = \frac{1}{4}x\Big]_{-2}^{1} = \frac{1}{4}x\Big]_{-2}^{1}$$

$$= \frac{1-(-2)}{2} = \frac{3}{4}.$$

$$P(|X| > 1) = 1 - P(|X| \le 1)$$

$$= 1 - P(-1 \le X \le 1)$$

$$= 1 - \int_{-1}^{1} f(x)dx = 1 - \frac{1}{4}x\Big]_{-1}^{1} = \frac{1}{2}$$

To find $P(2X + 3 > 5)$

$$P(2X + 3 > 5) = P(2X > 5 - 3)$$

$$= P(2X > 2)$$

$$= P(X > 1)$$

$$= 1 - P(X < 1) = 1 - \frac{3}{4} = 1/4$$

MOMENTS

If X is a random variable which is discrete or continuous, the moments about the origin denoted by μ_r' is defined as

$\mu_r' = E(X^r)$ for r = 1,2,3,…

The moments about the mean or central moments denoted by μ_r is defined as

$\mu_r = E\left[\left(X - \overline{X}\right)^r\right]$, for r=1,2,3,…

If X is a discrete random variable which can assume any one of the values $x_1, x_2,\ldots,x_n$ with respective probabilities $p(x_1),p(x_2),\ldots,p(x_n)$, then

$$\mu_r' = E(X^r) = \sum_{r=1}^{\infty} x^r p(x)$$

and $\mu_r = E\left[\{X - \overline{X})^r\right] = \sum (x - \bar{x})^r p(x_r)$.

If X is a continuous random variable with PDF f(x) then

$$\mu_r' = \int_{-\infty}^{\infty} x^r f(x)dx, \text{ r} = 1,2,3,..$$

and $\mu_r = \int_{-\infty}^{\infty} (x - \bar{x})^r f(x)dx, \text{ r} = 1,2,3,..$

Relation between moments about origin and moments about mean $\overline{X}$

By definition, $\mu_r = \left[E(X - \overline{X})^r \right]$

$$= E\left(X^r - r_{c_1} X^{r-1}\overline{X} + r_{c_2} X^{r-2}\overline{X}^2 - r_{c_3} X^{r-3}\overline{X}^3 + ... + (-1)^{r-1} r_{c_{r-1}} X\overline{X}^{r-1} + (-1)^r \overline{X}^r \right)$$

$$= E(X^r) - rE(X^{r-1})\overline{X} + r_{c_2} E(X^{r-2})\overline{X}^2 - r_{c_3} E(X^{r-3})\overline{X}^3 + ... + (-1)^r \overline{X}^r$$

Since $E(X) = \overline{X} = \mu_1'$ we have

$$\mu_r = \mu_r' - r\mu_{r-1}'\mu_1' + r_{c_2}\mu_{r-2}'\mu_1'^2 - r_{c_3}\mu_{r-3}'\mu_1'^3 + ... + (-1)^r \mu_1'^r$$

Results

1. The first moment about the mean is always zero, since $\mu_1 = \mu_1' - \mu_0'\mu_1' = 0$
2. The first moment about the origin is mean.
3. The second moment about the mean

$$\mu_2 = \mu_2' - 2\mu_1'\mu_1' + \mu_1'^2$$

$$\mu_2 = \mu_2' - \mu_1'^2$$

Var (X)= $E(X^2) - [E(X)]^2$

i.e., The second moment about the mean is variance.

4. The third moment about the mean is

$$\mu_3 = \left[E(X - \overline{X})^3 \right] = \mu_3' - 3\mu_2'\mu_1' + 3_{c_2}\mu_1'(\mu_1')^2 - 3_{c_3}(\mu_1')^3$$

$$= \mu_3' - 3\mu_2'\mu_1' + 3(\mu_1')^3 - \mu_1'^3$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^3$$

Similarly,

$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6(\mu_1')^2(\mu_2') - 3\mu_1'^4$ and so on.

Relation between moments about any point A and moments about mean $\overline{X}$.

W.K.T. $\mu_r' = E\left[(X - A)^r\right]$

Putting r =1,

$$\mu_1' = E[(X - A)] = \overline{X} - A \Rightarrow \text{Mean } \overline{X} = \mu_1' + A$$

Putting r = 2,

$$\mu_2' = \mu_2 + (\mu_1')^2$$

Similarly we get,

$$\mu_3' = \mu_3 + 3\mu_2\mu_1' + (\mu_1')^3$$

$$\mu_4' = \mu_4 + 4\mu_3\mu_1' + 6\mu_2(\mu_1')^2 + (\mu_1')^4 \text{ etc.}$$

Properties of Moments

1. If X is a random variable the

$$E(aX + b) = aE(X) + b$$

Note: $E(X \pm Y) = E(X) + E(Y)$

2. If X is a random variable then $\text{Var}(aX + b) = a^2 \text{ var}(X)$

3. If X and Y are independent then $\text{Var}(\text{aX} \pm \text{bY}) = \text{a}^2\text{Var}(\text{X}) \pm b^2\text{Var}(\text{Y})$

4. If X and Y are independent then $E(XY) = E(X)E(Y)$

5. If X and Y are two random variables such that $Y \le X$ then $E(Y) \le E(X)$

Covariance (X,Y)

The covariance of two random variables is denoted by $\rho_{xy} = Cov(X, Y)$ which is defined as

Cov(X,Y)=E(XY)-E(X)E(Y)

Note:

i. If X and Y are independent random variables then Cov(X,Y)=0
ii. If X and Y are any two random variables then
iii. $\text{cov}(aX \pm bY) = a^2Var(X) + b^2Var(Y) \pm 2ab \text{ Cov(X, Y)}$
iv. $Cov(aX + b, CY + d) = ab\, Cov(X, Y)$

Various measures of Central tendency, dispersion, skewness and kurtosis for continuous probability distribution

Let f(x) be the P.D.F of a random variable X where X is defined from a to b. Then

Arithmetic Mean = $\int_a^b f(x)dx = \bar{x}$

Harmonic Mean $\frac{1}{H} = \int_a^b \frac{1}{x} f(x)dx$

Geometric Mean

Median: Median is the point which divides the entire distribution in two equal parts. In case of a continuous distribution median is the point which divides the total area into two parts. Thus if M is the median then

$$\int_a^M f(x)dx = \int_M^b f(x)dx = \frac{1}{2}$$

Mean deviation: **Mean deviation about the mean** μ_1' is given by

$$M.D. = \int_a^b |x - mean| f(x)dx$$

Mean deviation about an average A is given by

M.D. about A = $\int_a^b |x - A| f(x)dx$

Mode: Mode is the value of X for which f(x) is maximum.

i.e., mode is given by $f'(x) = 0$ and $f''(x) < 0$ if $x \in [a, b]$

Coefficient of skewness and kurtosis are defined as

$$\beta_1 = \frac{{\mu_3}^2}{{\mu_2}^3}; \beta_2 = \frac{\mu_4}{{\mu_2}^2}$$

MOMENT GENERATING FUNCTION

Definition

The moment generating function of a random variable X denoted by $M_X(t)$ is defined as

$$M_X(t) = E(e^{tx})$$

$$\therefore M_X(t) = E\left(1 + \frac{tx}{1!} + \frac{(tx)^2}{2!} + \frac{(tx)^3}{3!} + \ldots\right)$$

$$= E(1) + \frac{t}{1!}E(X) + \frac{t^2}{2!}E(X^2) + \frac{t^3}{3!}E(X^3) + \ldots + \frac{t^r}{r!}E(X^r) + \ldots$$

$$= 1 + \frac{t}{1!}\mu_1' + \frac{t^2}{2!}\mu_2' + \frac{t^3}{3!}\mu_3' + \ldots + \frac{t^r}{r!}\mu_r' + \ldots$$

$$= \sum_{r=0}^{\infty} \frac{t^r}{r!}\mu_r'$$

Which gives the MGF interms of moments.

$\therefore$ $\frac{t^r}{r!}$ $M_X(t)$ μ_r' $\mu_r' = E(X^r)$

The co efficient of $r!$ in of where r = 1,2,… and , moment about the origin.

Moment Generating Function of X about any point X=a is defined as

$$M_X(t) = E(e^{t(x-a)}) = E\left(1 + \frac{t(x-a)}{1!} + \frac{t(x-a)^2}{2!} + \frac{t(x-a)^3}{3!} + \ldots\right)$$

$= 1 + \frac{t}{1!}\mu_1' + \frac{t^2}{2!}\mu_2' + \frac{t^3}{3!}\mu_3' + \ldots$ where $E\left((X-a)^r\right) = \mu_r'$.

Since $M_x(t)$ generates moments it is called moment generating function.

If X is a discrete random variable with PMF p(x) then

$$M_X(t) = E(e^{tx}) = \sum_x e^{tx} p(x)$$

If X is a continuous random variable with PDF f(x) then

$$M_X(t) = E(e^{tx}) = \int_{-\infty}^{\infty} e^{tx} f(x)dx$$

Moments using Moment Generating Function

Differentiate equation with respect to t and then putting t = 0 gives

$$\mu_1' = \left[\frac{d}{dt}M_X(t)\right]_{t=0}$$

$$\mu_2' = \left[\frac{d^2}{dt^2} M_X(t)\right]_{t=0}$$

In general,

$$\mu_r' = \left[\frac{d^r}{dt^r} M_X(t)\right]_{t=0} \text{ r=1,2,3,…}$$

Note:

1. Moment generating function $M_X(t)$ is used to calculate the higher moments.

2. $M_{aX}(t) = M_X(at)$, a being a constant.

3. The moment generating function of the sum of a given number of independent random variables is equal to the product of their respective moment generating function

i.e., $M_{X_1+X_2+...+X_n}(t) = M_{X_1}(t).M_{X_2}(t)...M_{X_n}(t)$.

4. Mean = μ_1'

5.Variance = $\mu_2' - \mu_1'^2$

Exercises

1) Find the M.G.F. for the following function given by

X	0	1	2	3	4	5	6
P(X)	1/49	3/49	5/49	7/49	9/49	11/49	13/49

Solution

$$M_X(t) = \sum_{x=0}^{t} e^{tx} p(x)$$

$$= \sum_{x=0}^{t} e^{tx} p(x)$$

$$= e^0(1/49) + e^t(3/49) + e^{2t}(5/49) + e^{3t}(7/49) + e^{4t}(9/49) + e^{5t}(11/49) + e^{6t}(13/49)$$

$$= \frac{1}{49}\left[1 + 3e^t + 5e^{2t} + 7e^{3t} + 9e^{4t} + 11e^{5t} + 15e^{6t}\right]$$

2) A random variable X has the probability function

$f(x) = \frac{1}{2^x}$, x =1,2,3,…

Find its i) M.G.F. ii) Mean

Solution

It is a discrete random variable ,

$$M.G.F. = \sum_{x=1}^{\infty} e^{tx} p(x)$$

$$= \sum_{x=1}^{\infty} e^{tx} f(x)$$

$$= \sum_{x=1}^{\infty} e^{tx} \frac{1}{2^x}$$

$$= \sum_{x=1}^{\infty} \left(\frac{e^t}{2}\right)^x$$

$$= \frac{e^t}{2} + \left(\frac{e^t}{2}\right)^2 + \left(\frac{e^t}{2}\right)^3 + \ldots$$

$$= \frac{e^t}{2}\left(1 + \frac{e^t}{2} + \left(\frac{e^t}{2}\right)^2 + \left(\frac{e^t}{2}\right)^3 + \ldots\right)$$

$$= \frac{e^t}{2}\left(1 - \frac{e^t}{2}\right)^{-1} = \frac{e^t}{2 - e^t}$$

$$\therefore M_X(t) = \frac{e^t}{2 - e^t}$$

Mean $$\overline{X} = \left[\frac{d}{dt}\left(\frac{e^t}{2 - e^t}\right)\right]_{t=0}$$

$$= \left[\frac{(2 - e^t)e^t - e^t(-e^t)}{(2 - e^t)^2}\right]_{t=0}$$

$$= \left[\frac{2 - e^{2t} + e^{2t}}{(2 - e^t)^2} \right]_{t=0} = \frac{2}{1} = 2 .$$

$$\text{variance} = \mu_2' - (\mu_1')^2$$

$$\mu_2' = \left[\frac{d^2}{dt^2} \left(\frac{e^t}{2 - e^t} \right) \right]_{t=0}$$

$$= \left[\frac{d}{dt} \left(\frac{2}{(2 - e^t)^2} \right) \right]_{t=0}$$

$$= \left[\frac{2(-2)(-e^t)}{(2 - e^t)^3} \right]_{t=0}$$

$$\mu_2' = 4 .$$

But $\text{variance} = \mu_2' - (\mu_1')^2$ =4-4=0.

3) Find the M.G.F for the distribution

f(x) = 2/3 at x = 1
= 1/3 at x = 2
= 0, otherwise

Also find i) Mean, ii) Variance

Solution

M.G. F. = $E(e^{tx})$

$$= \sum e^{tx} p(x)$$

$$M_X(t) = e^t 2/3 + e^{2t} .1/3 + 0$$

$$\mu_1' = \left[\frac{d}{dt} M_X(t) \right]_{t=0}$$

$$\mu_1' = \left[\frac{d}{dt} (2e^t / 3 + e^{2t} / 3 \right]_{t=0}$$

$$\mu_1' = \left[(2e^t / 3 + 2 e^{2t} / 3 \right]_{t=0}$$

$$= \frac{2}{3} + \frac{2}{3} = \frac{4}{3}$$

Variance = $\mu_2' - \mu_1'^2$

$$\mu_2' = \left[\frac{d^2}{dt^2} M_X(t) \right]_{t=0}$$

$$\mu_2' = \frac{d}{dt} \left\{ \frac{2e^t}{3} + \frac{2e^{2t}}{3} \right\}_{t=0}$$

$$\mu_2' = \left\{ \frac{2e^t}{3} + \frac{4e^{2t}}{3} \right\}_{t=0}$$

$$= \frac{2}{3} + \frac{4}{3} = \frac{6}{3} = 2$$

Variance= $\mu_2' - \mu_1'^2 = 6/3 - (4/3)^2$

$$= \frac{6}{3} - \frac{16}{9} = \frac{18-16}{9} = 2/9.$$

4) The moment generating function of a random variable X is given by

$$M_X(t) = \frac{e^t}{3} + \frac{4e^{3t}}{15} + \frac{2e^{4t}}{15} + \frac{4e^{5t}}{15}$$

Find the probability density function of X.

Solution

For a discrete random variable, $M_X(t) = E(e^{tx}) = \sum e^{tx} p(x)$

Hence the probability function is given by

x	1	3	4	5
P(x)	1/3	4/15	2/15	4/15

5) Find the M.G.F. of the random variable whose moments are $\mu_r' = (r+1)3^r$ and hence find its mean.

Solution

The M.G.F. is $M_X(t) = \sum_{r=0}^{\infty} \frac{t^r}{r!} \mu_r'$.

$$= \sum_{r=0}^{\infty} \frac{t^r}{r!} (r+1)! 3^r$$

$$= \sum_{r=0}^{\infty} \frac{(3t)^r}{r!} (r+1)!$$

$$= \sum_{r=0}^{\infty} (3t)^r (r+1)$$

$$M_X(t) = 1 + 2(3t) + 3(3t)^2 + \ldots$$

$$= (1-3t)^{-2}$$

$$M_X(t) = \frac{1}{(1-3t)^2}$$

$$\text{Mean} = \left[\frac{\mathrm{d}}{\mathrm{dt}} \{M_X(t)\} \right]_{t=0}$$

$$= \left[(-3) \frac{(-2)}{(1-3t)^3} \right]_{t=0}$$

$$= \left[\frac{6}{(1-3t)^3} \right]_{t=0}.$$

Mean = 6

(Or) $\text{Mean} = \mu_1' = \text{coefficient of } \frac{\mathrm{t}}{1!} = 6$.

6) Find the moment generating function of the distribution

$$\mathrm{f(x)} = \mathrm{ke}^{-\mathrm{kx}}, \mathrm{x} > 0$$
$$= 0, \text{otherwise}$$

Hence find i) Mean and ii) Variance, iii) μ_3' , iv) μ_4' .

Solution

Given

$f(x) = ke^{-kx}, x > 0$

$= 0, \text{otherwise}$

$\text{M.G.F.} = M_X(t) = E(e^{tx})$

$$= \int_{-\infty}^{\infty} e^{tx} f(x) dx$$

$$= \int_{-\infty}^{\infty} e^{tx} k e^{-kx} dx$$

$$= k \int_{0}^{\infty} e^{(t-k)x} dx$$

$$= k \frac{e^{(t-k)x}}{t-k} \Bigg]_0^{\infty}$$

$$= \frac{-k}{k-t}(0-1)$$

$$\therefore M_X(t) = \frac{k}{k-t}$$

$$\text{Mean} = \mu_1' = \left[\frac{d}{dt} M_X(t) \right]_{t=0}$$

$$= \left[\frac{d}{dt} \left(\frac{k}{k-t} \right) \right]_{t=0}$$

$$= \left[\frac{k(-(-1))}{(k-t)^2} \right]_{t=0}$$

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

Next to find Variance

Variance= $\mu_2' - (\mu_1')^2$

$$\mu_2' = \left[\frac{d^2}{dt^2} M_X(t)\right]_{t=0}$$

$$= \left[\frac{d}{dt}\left(\frac{k}{(k-t)^2}\right)\right]_{t=0}$$

$$= \left[\frac{-k.2(-1)}{(k-t)^3}\right]_{t=0}$$

$$= \left[\frac{2k}{(k-t)^3}\right]_{t=0} = \frac{2k}{k^3} = \frac{2}{k^2}$$

Variance= $\mu_2' - (\mu_1')^2$

$$= \frac{2}{k^2} - \left(\frac{1}{k}\right)^2 = \frac{1}{k^2}$$

$$\mu_3' = \left[\frac{d^3}{dt^3} M_X(t)\right]_{t=0}$$

$$\mu_3' = \left[\frac{d}{dt}\left(\frac{2k}{(k-t)^3}\right)\right]_{t=0}$$

$$\mu_3' = \left[\frac{2k(-3)(-1)}{(k-t)^4}\right]_{t=0}$$

$$= \frac{6k}{k^4} = \frac{6}{k^3}$$

$$\mu_4' = \left[\frac{d^4}{dt^4} M_X(t)\right]_{t=0}$$

$$\mu_4' = \left[\frac{d}{dt}\left(\frac{6k}{(k-t)^4}\right)\right]_{t=0}$$

$$\mu_4' = \left[\frac{6k(-4)(-1)}{(k-t)^5}\right]_{t=0}$$

$$' = \left[\frac{24k}{(k-t)^5} \right]_{t=0}$$

$$\frac{24k}{k^5} = \frac{24}{k^4}.$$

7) If the random variable X has the M.G.F.

$$M_X(t) = \frac{3}{3-t} = \frac{3}{3(1-t/3)}$$

$$= (1 - \frac{t}{3})^{-1}$$

$$= 1 + t/3 + (t/3)^2 + (t/3)^3 + ...$$

$$= 1 + \frac{t}{1!}(1/3) + \frac{t^2}{2!}(2/9) + \frac{t^3}{3!}(6/27) + ...$$

$$\mu_r' = \text{co} - \text{efficient of } \frac{\text{t}^\text{r}}{r!}$$

$$\mu_1' = \text{co} - \text{efficient of } \frac{\text{t}}{1!} = 1/3$$

$$\mu_2' = \text{co} - \text{efficient of } \frac{\text{t}^2}{2!} = 2/9$$

Variance = $\mu_2' - (\mu_1')^2$

$$= \frac{2}{9} - (\frac{1}{3})^2 = \frac{1}{9}$$

Standard deviation = $\sqrt{\text{Variance}} = \frac{1}{3}$.

8) Find the m.g.f. of the distribution defined by $dF = \frac{1}{2} e^{-|x|} dx$, $-\infty < x < \infty$ and hence find the variance.

Solution

$$dF = \frac{1}{2} e^{-|x|} dx$$

$\frac{dF}{dx} = \frac{1}{2} e^{-|x|}$.

Now $\frac{dF}{dx} = f(x)$. i.e., $f(x) = \frac{1}{2} e^{-|x|}$.

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} f(x) dx$$

$$= \int_{-\infty}^{\infty} e^{tx} \frac{1}{2} e^{-|x|} dx$$

$$= \int_{-\infty}^{0} e^{tx} \frac{1}{2} e^{-(-x)} dx + \int_{0}^{\infty} e^{tx} . \frac{1}{2} e^{-x} dx$$.

$$= \frac{1}{2} \int_{-\infty}^{0} e^{(t+1)x} dx + \frac{1}{2} \int_{0}^{\infty} e^{t(-1)x} dx$$

$$= \frac{1}{2} \int_{-\infty}^{0} e^{(t+1)x} dx + \frac{1}{2} \int_{0}^{\infty} e^{(t-1)x} dx$$

$$= \frac{1}{2} \frac{e^{(t+1)x}}{t+1} \Big]_{-\infty}^{0} + \frac{1}{2} \left(\frac{-e^{-(1-t)x}}{1-t} \right)_{0}^{\infty}$$

$$= \frac{1}{2} \left[\frac{1}{1+t} - 0 + \frac{-1}{1-t} (0-1) \right]$$

$$= \frac{2}{2(1-t^2)} = \left(1-t^2\right)^{-1} = 1 + t^2 + \frac{(t^2)^2}{2!} + \frac{\left(t^2\right)^3}{3!} + \ldots$$

$\therefore \mu_1' = \text{co efficient of } t/1! = 0$

$\mu_2' = \text{co efficient of } t^2/2! = 2$

Variance = $\mu_2' - (\mu_1')^2 = 2 - 0^2 = 2$

UNIT-II

PROBABILITY DISTRIBUTIONS

Introduction

While constructing probabilistic models for observable phenomena, certain probability distributions arise more frequently than do others. we treat such distributions that play important roles in many engineering applications as special probability distributions.

DISCRETE DISTRIBUTIONS

Bernoulli Trials and Bernoulli Distributions

Let A be an event ((trail) associated with a random experiment such that p(A) remains the same for the repetitions of that random experiment, then the events are called Bernoulli trails.

A random variable X which takes only two values either 1 (success) or 0(failure) with probability p and q respectively. i.e., P(X=1)=p, P(X=0)=q, p+q=1 is called Bernoulli variate and is said to have a Bernoulli distribution.

Moments of Bernoulli Distribution

$$\mu_{r'} = E(X^r) = 1^{r\cdot} \cdot p + 0^r \cdot q = p$$

$$\mu_1' = E(X) = p, \mu_2' = E(X^2) = p$$

Mean=p

$$Var(X) = E(X^2) - [E(X)]^2 = p - p^2 = p(1-p) = pq$$

Definition.

A random variable X is said to follow binomial distribution denoted by B(n,p) if it assumes only non-negative values and its probability mass function is given by

$$p(x) = P(X = x) = n_{c_x} p^x q^{n-x}$$,x=0,1,2,…,n

=0, otherwise

Where n and p are parameters.

Binomial Frequency Distribution

Suppose that n trails constitute an experiment and if this experiment is repeated N times the frequency function of the binomial distribution is given by

$Np(x) = N \times n_{c_x} p^x q^{n-x}$, x = 0,1,2,...,n

Properties of Binomial Frequency Distribution

1. Each trail results in two mutually disjoint outcomes, termed success and failure.

2. The trails must be independent of each other.

3. All trails have same constant probability of success.

4. The number of trails n is finite.

Mean of Binomial Distributions

Mean = $E(X) = \sum_x xp(x)$

$= \sum_{x=0}^{n} x n_{c_x} p^x q^{n-x}$

$= \sum_{x=0}^{n} x \cdot \frac{n!}{x!(n-x)!} p^x q^{n-x}$

$= \sum_{x=0}^{n} \frac{n(n-1)!\, p p^{x-1} q^{n-x}}{(x-1)!(n-x)!}$

$= np \sum_{x=1}^{n} \frac{(n-1)!\, p^{x-1} q^{n-x}}{(x-1)!(n-x)!}$

$= np \sum_{x=1}^{n} \frac{(n-1)!\, p^{x-1} q^{n-x}}{(x-1)!(n-x)!}$

$= np \sum_{x=1}^{n} (n-1)_{c_{x-1}} p^{x-1} q^{n-x}$

$= np \sum_{x=1}^{n} (n-1)_{c_{x-1}} p^{x-1} q^{(n-1)-(x-1)}$

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

Mean=np

Variance of Binomial Distribution

$Var(X) = E(X^2) - [E(X)]^2$

The probability mass function of binomial distribution is

$P(X = x) = p(x) = n_{c_x} p^x q^{n-x}$, x = 0,1,2,...,n

$$E(X^2) = \sum_{x=0}^{n} x^2 p(x)$$

$$= \sum_{x=0}^{n} x^2 n_{c_x} p^x q^{n-x}$$

$$= \sum_{x=0}^{n} x^2 \frac{n!}{x!(n-x)!} n_{c_x} p^x q^{n-x}$$

$$= \sum_{x=0}^{n} [x(x-1)+x] \frac{n!}{x!(n-x)!} p^x q^{n-x}$$

$$= \sum_{x=0}^{n} x(x-1) \frac{n!}{x!(n-x)!} p^x q^{n-x} + \sum_{x=0}^{n} x \frac{n!}{x!(n-x)!} p^x q^{n-x}$$

$$= \sum_{x=0}^{n} \frac{n(n-1)(n-2)!}{(x-2)!(n-x)!} p^2 p^{x-2} q^{n-x} + E(X)$$

$$= n(n-1)p^2 \sum_{x=0}^{n} \frac{(n-2)!}{(x-2)!(n-x)!} p^{x-2} q^{n-x} + np$$

$$= n(n-1)p^2 (q+p)^{n-2} + np$$

$$E(x^2) = n(n-1)p^2 + np$$

But, $\mathrm{Var(X)} = \mathrm{E(X^2)} - [E(X)]^2$

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

$= p^2\left[n^2 - n - n^2\right] + np$

$= np(1-p)$

$= npq$

Note

Similarly we can prove that

i. $\mu_3 = npq(q-p)$

ii. $\mu_4 = npq(1-6pq+3npq)$

iii. Mean of the binomial distribution always greater than its variance.

Moment Generating Function (M.G.F)

The probability mass function of a binomial distribution is

$$P(X=x) = n_{c_x} p^x q^{n-x}, x = 0,1,2,...n$$

where n is the number of independent trials and x is the number of success.

By definition of the moment generating function

$M_x(t) = E(e^{tx})$

$$= \sum_{x=0}^{n} e^{tx} n_{c_x} p^x q^{n-x}$$

$$= \sum_{x=0}^{n} n_{c_x} (pe^t)^x q^{n-x}$$

$$= q^n + n_{c_1}(pe^t)^1 q^{n-1} + n_{c_2}(pe^t)^2 q^{n-2} + ...+ (pe^t)^n$$

$$= (q + pe^t)^n$$

Finding mean using moment generating function:

$$E(X) = \mu_1' = \left[\frac{d}{dt} M_X(t)\right]_{t=0}$$

$$= \left[\frac{d}{dt}(q + pe^t)^n\right]_{t=0}$$

$$= \left[n(q + pe^t)^{n-1} pe^t\right]_{t=0}$$

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

Finding variance using moment generating function

$$E(X^2) = \mu_2' = \left[\frac{d^2}{dt^2} M_X(t)\right]_{t=0}$$

$$= \left[\frac{d^2}{dt^2}(q + pe^t)^n\right]_{t=0}$$

$$= \left[npe^t(q + pe^t)^{n-1} + n(n-1)(q + pe^t)^{n-2}(pe^t)^2\right]_{t=0}$$

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

$$= \left[np + n(n-1)p^2\right]$$

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

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

$$E(X^2) = npq + n^2 p^2$$

$$Var(X) = E(X^2) - [E(X)]^2$$

$$= npq + n^2 p^2 - n^2 p^2$$

$$= npq$$

Moment Generating Function of Binomial Distribution about mean

By definition,

$$M_{X-np}(t) = E[e^{t(x-np)}]$$

$$= E[e^{tx-tnp}]$$

$$= E[e^{tx}e^{-tnp}]$$

$$= e^{-npt}E[e^{tx}]$$

$$M_X(t) = e^{-npt}(q + pe^t)^n$$

Additive or Reproductive property of binomial distribution

If X_1 and X_2 are two independent binomial variates with parameters (n_1,p) and (n_2,p) respectively then X_1+X_2 is a binomial variate with parameter (n_1+n_2,p)

Proof

The MGF of the random variable X_1+X_2 is

$$M_{X_1+X_2}(t) = M_{X1}(t)M_{X_2}(t)$$

$$= (q + pe^t)^{n_1}(q + pe^t)^{n_2}$$

$$= (q + pe^t)^{n_1+n_2}$$

This shows that X_1+X_2 is also a binomial variate with parameters n_1+n_2 and P.

Note

If X_1 and X_2 are two independent binomial variates with parameters (n_1,p_1) and (n_2,p_2) then X_1+X_2 is not a binomial variate.

Recurrence formula for central moment

$$\mu_k = E[X - E(X)]^k = \sum_{x=0}^{n}(x - np)^k n_{c_x} p^x q^{n-x} \quad \text{----------(i)}$$

Differentiate with respect to P we get

$$\frac{d}{dp}\mu_k = \sum_{x=0}^{n} n_{c_x}\left\{-nk(x-np)^{k-1}p^x q^{n-x}\right\} + (x-np)^k[xp^{x-1}q^{n-x} + p^x(n-x)q^{n-x-1}(-1)]$$

$$= -nk\mu_{k-1} + \sum_{x=0}^{n} n_{c_x} (x-np)^k \, p^{x-1} q^{n-x-1} [xq - (n-x)p]$$

$$= -nk\mu_{k-1} + \frac{1}{pq} \sum_{x=0}^{n} n_{c_x} p^x q^{n-x} (x-np)^k (x-np) \qquad Since, (p+q=1)$$

$$= -nk\mu_{k-1} + \frac{1}{pq} \sum_{x=0}^{n} n_{c_x} p^x q^{n-x} (x-np)^{k+1}$$

$$\frac{d\mu_k}{dp} = -nk\mu_{k-1} + \frac{1}{pq} \mu_{k+1}$$

Therefore, $\mu_{k+1} = pq\left\{\frac{d\mu_k}{dp} + nk\mu_{k-1}\right\}$ ----------(ii)

Using recurrence relation (ii) we can compute moments of higher order, provided the moment of lower order are known.

Putting k=1 in (ii)

$$\mu_2 = pq\left\{\frac{d\mu_1}{dp} + n\mu_0\right\} = npq \qquad Since, (\mu_0 = 1, \mu_1 = 0)$$

Therefore, $\mu_2 = npq$

Putting k=2 in (ii)

$$\mu_3 = pq\left\{\frac{d\mu_2}{dp} + 2n\mu_1\right\}$$

$$= pq\left\{\frac{d}{dp}(npq) + 2n(0)\right\}$$

$$= pq\left\{\frac{d}{dp}[np(1-p)]\right\} \qquad Since, q = 1-p$$

$$= pq\left\{\frac{d}{dp}[np - np^2]\right\}$$

$$= pq\{n - 2np\}$$

$$= npq - 2np^2q$$

$$\mu_3 = npq(1-2p)$$

Putting k=3 in (ii)

$$\mu_4 = pq\left\{\frac{d\mu_3}{dp} + 3n\mu_2\right\}$$

$$= pq\left\{\frac{d}{dp}\left[npq(1-2p)\right] + 3n(npq)\right\}$$

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

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

$$= pq\left\{\frac{d}{dp}\left[np - 2np^2 - np^2 + 2np^3\right] + 3n^2pq\right\}$$

$$= pq\left\{n - 4np \quad - 2np \quad + 6np^2 + 3n^2pq\right\}$$

$$= pq\left\{n - 6np + 6np^2 + 3n^2pq\right\}$$

$$= npq\left\{1 - 6p + 6p^2 + 3npq\right\}$$

$$= npq\{1 - 6p + 6p(1-q) + 3npq\}$$

$$= npq\{1 - 6pq + 3npq\}$$

$$= npq\{1 + 3pq(n-2)\}$$

Note

μ_2 is the variance

μ_3 is a measure of skewness and

μ_4 is a measure of kurtosis

We denote sometime the measure of skewness and kurotsis by β_1 and β_2

$$\beta_1 = \frac{{\mu_3}^2}{{\mu_2}^3} \quad , \quad \beta_2 = \frac{\mu_4}{{\mu_2}^2}$$

Examples

1. The mean and variance of a binomial distribution are 4 and 4/3 respectively. Find P(X≥1) if n=6.

Solution

Mean of binomial distribution = np = 4

Variance of binomial distribution = npq = 4/3

$$\frac{npq}{np} = \frac{\frac{4}{3}}{4}$$

$$q = 1/3$$

Now p = 1-q = 1-1/3 = 2/3

Given n=6

$$P(X = x) = n_{c_x} p^x q^{n-x}$$

$$P(X \geq 1) = 1 - P[X < 1]$$

$$= 1 - P[X = 0]$$

$$= 1 - 6_{c_0} p^0 q^{6-0}$$

$$= 1 - q^6$$

$$= 1 - (\frac{1}{3})^6$$

$$= 1 - \frac{1}{729}$$

$$= \frac{728}{729}$$

2. The mean and variance of binomial distributions are 4 and 3 respectively. Find P(X=0), P(X=1) and P(X≥2).

Solution

Mean of binomial distribution = np = 4

Variance of binomial distribution = npq= 3

$$\frac{npq}{np} = \frac{3}{4}$$

$$q = \frac{3}{4}$$

Now p = 1-q = 1-3/4 = 1/4

Since Mean = np = 4

= n(1/4) = 4

n = 16

$$P(X = x) = n_{c_x} p^x q^{n-x}$$

$$P(X = 0) = n_{c_0} p^0 q^n$$

$$= 16_{c_0} (\frac{3}{4})^{16}$$

$$= (\frac{3}{4})^{16} = 0.01$$

$$P(X = 1) = n_{c_1} p^1 q^{n-1}$$

$$= 16_{c_1} p^1 q^{15}$$

$$= 16(\frac{1}{4})(\frac{3}{4})^{15} = 0.053$$

$$P(X \geq 2) = 1 - P(X < 2)$$

$$= 1 - [P(X = 0) + P(X = 1)]$$

$$= 1 - [0.01 + 0.053] = 1 - 0.063$$

$$= 0.937$$

3. If the mean is 3 and variance is 4 of a random variable X, check whether X follows binomial distribution,

Solution

No. Because for a binomial distribution mean should be greater than the variance.

If mean = np = 3 and variance = npq = 4

npq/np = q = 4/3 = 1.33

1.33 is greater than 1

q>1 (but the probability is less than 1)

Therefore mean should be greater than the variance for a binomial distribution.

3. A binomial variate X satisfies the relation 9P(X=4) = P(X=2) when n=6. Find the parameter p of the binomial distribution.

Solution

The probability function for a binomial distribution is

$$P(X = x) = n_{c_x} p^x q^{n-x}$$

$$P(X = 4) = 6_{c_4} p^4 q^{6-4}$$

$$P(X = 4) = 6_{c_4} p^4 q^2$$

$$P(X = 2) = 6_{c_2} p^2 q^4$$

Given 9P(X=4) = P(X=2)

$$9 * 6_{c_4}\, p^4 q^2 = 6_{c_2}\, p^2 q^4$$

$$135 p^2 = 15 q^2$$

$$9 p^2 = q^2$$

$$9 p^2 - q^2 = 0$$

$$9 p^2 - (1 - p)^2 = 0$$

$$9 p^2 - (1 + p^2 - 2p) = 0$$

$$9 p^2 - 1 - p^2 + 2p = 0$$

$$8 p^2 + 2p - 1 = 0$$

$$p = \frac{-2 \pm \sqrt{4 + 32}}{16}$$

$$p = \frac{-2 \pm 6}{16} \qquad = \frac{4}{16}, \frac{-8}{16}$$

$$p = \frac{1}{4}, \frac{-1}{2}$$

Since p cannot be negative, p=1/4.

4. Out of 800 families with 4 children each, how many families would be expected to have

(i) 2 boys and 2 girls

(ii) at least 1 boy

(iii) at most 2 girls and

(iv) children of both sexes.

Assume equal probabilities for boys and girls.

Solution

Considering each child is a trial, n=4. Assuming that birth of a boy is success, p = 1/2 and q = ½

Let X denote the number of successes (boys)

(i) P[2 boys and 2 girls] = P(X=2)

$$P(X = x) = {}_{n}c_{x}\, p^{x} q^{n-x}$$

$$P(X = 2) = {}_{4}c_{2} \left(\frac{1}{2}\right)^{2} \left(\frac{1}{2}\right)^{4-2}$$

$$= 6\left(\frac{1}{2}\right)^{4} = \frac{3}{8}$$

Therefore number of families having 2 boys and 2 girls=N[P(X=2)]

= 800(3/8) = 100 * 3

= 300

(ii) P[at least 1 boy] = P[X≥1]

= P[X=1] + P[X=2] + P[X=3] + P[X=4]

=1- P[X=0]

$$P(X = 0) = {}_{4}c_{0} \left(\frac{1}{2}\right)^{0} \left(\frac{1}{2}\right)^{4-0}$$

$$1 - P(X = 0) = 1 - \left(\frac{1}{2}\right)^{4} = \frac{15}{16}$$

Therefore number of families having at least 1 boy = N [1-(P(X=0)]

= 800 (15/16) = 750

(iii) P(at most 2 girls)= P(exactly 0 girl, 1 girl or 2 girls)

= P[X=4, X=3, X=2]

= 1-[P(X=0) + P(X=1)]

$$= 1 - \left[{}_{4}c_{0} \left(\frac{1}{2}\right)^{0} \left(\frac{1}{2}\right)^{4-0} + {}_{4}c_{1} \left(\frac{1}{2}\right)^{1} \left(\frac{1}{2}\right)^{4-1} \right]$$

$$= 1 - \left[(\frac{1}{2})^4 + 4(\frac{1}{2})^4 \right] = 1 - (\frac{1}{16} + \frac{4}{16}) = 1 - \frac{5}{16}$$

$$= \frac{11}{16}$$

Therefore number of families having at most 2 girls = N[P(X≥2)]

= 800 (11/16) = 550

(iv) P[children of both sexes] = 1 – P[children of same sex]

= 1 –[P(all are boys) + P(all are girls)]

= 1- [P(X=4) + P(X=0)]

$$= 1 - \left[4_{c_4} (\frac{1}{2})^4 + 4_{c_0} (\frac{1}{2})^0 (\frac{1}{2})^4 \right] = 1 - [(\frac{1}{2})^4 + (\frac{1}{2})^4]$$

= 1- 2/16 = 7/8

Therefore number of families having children of both sexes = 800 * 7/8

= 700

5. An irregular 6 faced die is such that the probability that it gives 3 even numbers in 5 throws is twice the probability that it gives 2 even numbers in 5 throws. How many sets of exactly 5 trials can be expected to give no even number out of 2500 sets.

Solution

Let the probability of getting an even number with the unfair die be p .

Let X denote the number of even numbers obtained in 5 trials (throws)

Given: P(X=3) = 2 * P(X=2)

$$5_{c_3} p^3 q^2 = 2 * 5_{c_2} p^2 q^3$$

p = 2q

p = 2(1-p)

3p = 2

P = 2/3

q=1-p = 1/3

Now P[getting no even number] = P[X=0]

$$5_{c_0}\, p^0 q^5 = (\frac{1}{3})^5 = \frac{1}{243}$$

Therefore number of sets having no success (even number) out of N sets = N [P(X=0)]

= 2500 * 1/243

= 10 nearly

6. A communication system consists of n components each of which will independently function with probability P. The total system will be able to operate effectively is at least one half of is components function. For what values of p is a 5 component system more likely to operate effectively than a 3 component system?

Solution

Since p is a constant and the n components function independently, the number of components X that function follows a binomial distribution

$$p(x) = n_{c_x}\, p^x q^{n-x}, x = 0,1,2,..,n$$

P[5 component system functioning effectively] = P[X=3 or 4 or 5 }

= P[X=3] + P[X=4] + P[X=5]

$$= 5_{c_3}\, p^3 q^{5-3} + 5_{c_4}\, p^4 q^{5-4} + 5_{c_5}\, p^5 q^{5-5}$$

$$= 10 p^3 q^2 + 5p^4 + p^5$$

P[3 component system functioning effectively] = P(X≥2)

= P(X=2) + P(X=3)

$$= 3_{c_2}\, p^2 q^{3-2} + 3_{c_3}\, p^3 q^{3-3}$$

$$= 3p^2 q + p^3$$

5 component system will function more effectively than 3 component system,

$$10p^3q^2 + 5p^4 + p^5 \geq 3p^2q + p^3$$

$$p^2(10pq^2 + 5p^2q + p^3 - 3q - p) \geq 0$$

$$p^2(10p(1-p)^2 + 5p^2(1-p) + p^3 - 3(1-p) - p) \geq 0$$

$$p^2(10p(1 + p^2 - 2p) + 5p^2 - 5p^3 + p^3 - 3 + 3p - p) \geq 0$$

$$p^2(10p + 10p^3 - 20p^2 + 5p^2 - 5p^3 + p^3 - 3 + 2p) \geq 0$$

$$p^2(6p^3 - 15p^2 + 12p - 3) \geq 0$$

$$3p^2(2p^3 - 5p^2 + 4p - 1) \geq 0$$

$$3p^2[(2p^2 - 3p + 1)(p - 1)] \geq 0$$

$$3p^2[(p - 1/2)(p - 1)(p - 1)] \geq 0$$

$$3p^2(p - 1/2)(p - 1)^2 \geq 0$$

Since, $3p^2(p-1)^2 \geq 0$

We have p-1/2≥0

That is p≥1/2

7. At least one half of an airplane's engines are requires to function In order for it to operate. If each engine independently function with probability p, for what values of p is a 4 engine plane to be preferred for operatiions to a 2 engine plane?

Solution

For a 2 engine plane ,

P [airplane operates] = P(X≥1)

=P(X=1) + P(X=2)

$$= 2_{c_1} p^1 q^1 + p^2$$

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

$$= 2p - p^2 = p(2-p)$$

For a 4 engine plane,

P[airplane operates] = P(X≥2)

$$= 1- [P(X<2)$$

$$= 1- [P(X=0) + P(X=1)]$$

$$= 1 - [q^4 + 4_{c_1} p^1 q^3]$$

$$= 1 - q^2[q^2 + 4pq]$$

$$= 1 - (1-p)^2[(1-p)^2 + 4pq]$$

$$= 1 - (1-p)^2[(1 + p^2 - 2p) + 4p(1-p)]$$

$$= 1 - (1-p)^2[1 + p^2 - 2p + 4p - 4p^2]$$

$$= 1 - (1-p)^2[1 + 2p - 3p^2]$$

$$= 1 - (1 + p^2 - 2p)[1 + 2p - 3p^2]$$

$$= 1 - (1 + p^2 - 2p)(1) - (1 + p^2 - 2p)(2p) - (1 + p^2 - 2p)(-3p^2)]$$

$$= 1 - 1 - p^2 + 2p - 2p - 2p^3 + 4p^2 + 3p^2 + 3p^4 - 6p^3)$$

$$= 3p^4 - 8p^3 + 6p^2$$

$$= p^2(3p^2 - 8p + 6)$$

Since a 4 engine plane is preferred than a 2 engine plane,

$$p^2(3p^2 - 8p + 6) > p(2-p)$$

$$3p^4 - 8p^3 + 6p^2 > 2p - p^2$$

$$3p^4 - 8p^3 + 7p^2 - 2p > 0$$

$$p(3p^3 - 8p^2 + 7p - 2) > 0$$

$$(3p^3 - 8p^2 + 7p - 2) \geq 0$$

$$(p-1)(3p^2 - 5p + 2) \geq 0$$

Now, $3p^2 - 5p + 2 = 3p^2 - 3p - 2p + 2$

$= 3p(p-1) - 2(p-1)$

$= (3p-2)(p-1)$

$$(p-1)(3p-2)(p-1) \geq 0$$

$$(p-1)^2(3p-2) \geq 0$$

That is

$$(p-1)^2 \geq 0, (3p-2) \geq 0$$

$P \geq 1$ or $p \geq 2/3$

Since $p>2/3$ is the only pemitted value and also $P<1$, the required value of p is 2/3.

8. A factory has 10 machines which may need adjustment from time to time during the day. Three of these machines are old, each having a probability of 1/11 of needing adjustment during the day and 7 are new, having the corresponding probability of 1/21. Assuming that no machines needs adjustments twice on the same day, find the probabilities that on a particular day,

(i) just 2 old and no new machine need adjustment and

(ii) just 2 machines that need adjustment are of the same type

Solution

Let X_1 be a random variable which denotes the number of old machines need adjustment and X_2 be the random variable which denotes the number of new machines that need adjustments

Let p_1= Probability that an old machine needs adjustment

p_1= 1/11 q_1=1-p_1= 10/11

p_2= Probability that a new machine needs adjustment

p_2= 1/21 q_2=20/21

There are 3 old machines i.e., n=3

$$P(X_1 = x) = n_{c_x} p_1{}^x q_1{}^{n-x}$$

$$= 3_{c_x} (\frac{1}{11})^x (\frac{10}{11})^{3-x}, x = 0,1,2,3$$

There are 7 new machines i.e., n=7

$$P(X_2 = x) = n_{c_x} p_2{}^x q_2{}^{n-x}$$

$$= 7_{c_x} (\frac{1}{21})^x (\frac{20}{21})^{7-x}, x = 0,1,2,..,7$$

The random variables X_1 and X_2 are independent.

i) The probability that just 2 old machines and no new machines need adjustment is given by

$P_1(X_1=2 \cap X_2=0) = P(X_1=2)P(X_2=0)$

$$= 3_{c_2} (\frac{1}{11})^2 (\frac{10}{11}).7_{c_0} (\frac{1}{21})^0 (\frac{20}{21})^7$$

=0.016

ii) If just 2 machines need adjustment and they are of the same type can happen in the following two mutually exclusive ways:

a) 2 old and no new machine (or)

b) 2 new and no old machine

Therefore the required probability is

$=P(X_1=2 \cap X_2=0) + P(X_1=0 \cap X_2=2)$

$$= 0.016 + (\frac{10}{11})^3 7_{c_2} (\frac{1}{21})^2 (\frac{20}{21})^5$$

=0.016+0.028

=0.044

9.The probability of a man hitting a target is 1/4.

i) If he fires 7 times, what is the probability of his hitting the target at least twice?

ii) How many times must he fire so that the probability of his hitting the target at least once is greater than 2/3?

Solution

Let X be a random variable which denotes the number of hits.

Given: p=1/4 and q=3/4

$P(X = x) = p(x) = n_{c_x} p^x q^{n-x}$, x = 0,1,2,...,n

i) Given n=7

$$P(X = x) = 7_{c_x} (\frac{1}{4})^x (\frac{3}{4})^{7-x}, \text{ x} = 0,1,2,...,7$$

P(hitting at least twice i.e., X≥2) = 1-[P(X=0)+P(X=1)

$$= 1 - \left[7_{c_0} (\frac{1}{4})^0 (\frac{3}{4})^7 + 7_{c_1} (\frac{1}{4})^1 (\frac{3}{4})^6 \right]$$

$$= 1 - (\frac{3}{4})^6 [\frac{3}{4} + \frac{7}{4}]$$

$$= 1 - (\frac{3}{4})^6 [\frac{10}{4}]$$

$$= 0.551$$

ii) P(hitting at least once) = P(X≥1)

= 1 – P(X<1)

= 1- P(X=0)

$= 1-(3/4)^n \geq 2/3$ (Given)

$1-2/3 \geq (3/4)^n$

$1/3 \geq (3/4)^n$

By trial, When n=4 this condition is satisfied. Therefor he must fire 4 times to hit at least once with probability more than 2/3.

10. A set of 6 similar coins are tossed 640 times with the following results:

Number of heads	0	1	2	3	4	5	6
Frequency	7	64	140	210	132	75	12

Calculate the binomial frequencies on the assumption that the coins are symmetrical.

Solution

Let X denote the number of heads and let X follow binomial distribution with parameters (n,p)

Here n=6

To find p, we compute the mean of given frequency distribution and equate it to np(mean od the binomial distribution)

X	0	1	2	3	4	5	6
F	7	64	140	210	132	75	12
Fx	0	64	280	630	528	375	72

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{1949}{640}$$

6p= 1949/640

p= 0.5075 q=0.4925

The distribution of X is given by

$$NP(X = x) = 640.6_{c_x}(0.5075)^x(0.4925)^{6-x}$$

The expected frequencies are calculated as follows

X	px)	Np(x)	Expected frequency
0	$(0.4925)^6$	9.13308	9
1	$6(0.5075)(0.4925)^5$	56.467	56
2	$15(0.5075)^2(0.4925)^4$	145.4683	146
3	$20(0.5075)^3(0.4925)^3$	199.865	200
4	$15(0.5075)^4(0.4925)^2$	154.4642	154
5	$6(0.5075)^5(0.4925)$	63.6675	64
6	$(0.5075)^6$	10.947	11
		Total	640

11.Two dice are thrown 120 times. Find the average number of times in which the number on the first die exceeds the number on the second die.

Solution

The number on the first die exceeds that on the second die, in the following combinations:

(2,1);

(3,1) ,(3,2);

(4,1),(4,2),(4,3);

(5,1),(5,2),(5,3),(5,4);

(6,1),(6,2),(6,3),(6,4),(6,5)

Where the numbers in the parantheses represent the numbers in the first and second dice respectively.

P(success) = P(number in the first dice exceed the number in the second dice)

=15/36 =5/12

This probability remains the same in all the throws that are independent.

If X is the number of success, then X follows binomial distribution with parameter n=120 and p= 5/12

E(X)=np=120*5/12= 50

12. It is known that diskettes produced by a certain company are defective with a probability 0.01 independently of each other. The company markets diskettes in packages of 10 and offers a money guarantee that almost 1 of the 10 diskettes is defective. What proportion of diskettes are returned? If someone buy 3 diskettes, what is the probability that he will return exactly one of them?

Solution

Given p = 0.01

q = 1-p =0.99

n = 10

Let X be the random variable denoting the number of defective in a package.

Then

$$P(X = x) = n_{c_x} p^x q^{n-x}$$

The company must replace the packages only when it has more than 1 defective diskette

P[at most 1 diskette is defective] = P[X≤1]

$$= P(X=0) + P(X=1)$$

$$= 10_{c_0} (0.01)^0 (0.99)^{10} + 10_{c_1(} (0.01)^1 (0.99)^9$$

$$= 0.9044 + 0.0914$$

$$= 0.1858$$

P[a package will have to replace]= P[X>1]

$$= 1 – P[X≤1]$$

$$= 1 – 0.1858 = 0.8142$$

Therefore, 0.4% of the packages will have to replace. [or the proportion of packages sold to be replace is 4%]

If someone buys 3 diskettes,

P[a package will have to return] = 3 * 0.1842

= 2.4426

13. Assuming that half of the population is vegetarian and that 100 investigators each take 10 individuals to see whether they are vegetarians, how many would you expect to report that 3 people or less were vegetarians?

Solution

n=10, p=1/2, q=1/2

$$P(X = x) = n_{c_x} p^x q^{n-x}$$

$$= 10_{c_x} (\frac{1}{2})^x (\frac{1}{2})^{10-x}$$

$$= 10_{c_x} (\frac{1}{2})^{10}$$

$$P(X \le 3) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)$$

$$= 10_{c_0} (\frac{1}{2})^{10} + 10_{c_1} (\frac{1}{2})^{10} + 10_{c_2} (\frac{1}{2})^{10} + 10_{c_3} (\frac{1}{2})^{10}$$

$$= (\frac{1}{2})^{10}[1 + 10 + 45 + 120] \backslash$$

$$= (\frac{1}{2})^{10}[176] = \frac{176}{1024} = 0.1718$$

Among 100 investigators, the number of investigators who report that 3 or less were consumers

=100 * 0.1718

=17 investigators

14. A factory produces 10 articles daily. It may be assumed that there is a constant probability p= 0.1 of producing a defective article. Before the articles are stored, they are inspected and the defective

ones are set aside. Suppose that there is a constant probability r = 0.1, that a defective article is misclassified. If X denote the number of articles classified as defective at the end of a production day, find a) P(X=3) and b) P(X>3)

Solution

Let X be the random variable represented by the number of articles which are defective.

P[a defective article is classified as defective] = P(an article produced is defective) * P(it is classified as defective)

= 0.1 *0.9

p = 0.09

q = 1 – p = 0.91

n= 10

$$P(X = x) = n_{c_x} p^x q^{n-x}$$

$$= 10_{c_x} (0.09)^x (0.91)^{10-x}$$

$$P(X = 3) = 10_{c_3} (0.09)^3 (0.91)^7$$

$$= 0.0452$$

$$P(X > 3) = 1 - P(X \le 3)$$

$$= 1 - [P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)]$$

$$= 1 - [10_{c_0} (0.09)^0 (0.91)^{10} + 10_{c_1} (0.09)^1 (0.91)^9 + 10_{c_2} (0.09)^2 (0.91)^8 + 10_{c_3} (0.09)^3 (0.91)^7]$$

=0.0089

15. An irregular 6 faced die is thrown and the expectation that in 10 throws it will give 5 even numbers is twice the expectation that it will give 4 even numbers. How many times in 10,000 sets of 10 throws would you expect to give even number?

Solution

Let the random variable X denote the number of even numbers.

$$\text{P[getting x even numbers]} = P(X = x) = n_{c_x}\, p^x q^{n-x}, x = 0,1,2...,n$$

Given n=10

$$P(X = x) = 10_{c_x}\, p^x q^{10-x}, x = 0,1,2...,10$$

P[getting 5 even numbers] = 2 * P[getting 4 even numbers]

$$10_{c_5}\, p^5 q^5 = 2 * 10_{c_4}\, p^4 q^6$$

$$\frac{6}{5} p^5 q^5 = 2 p^4 q^6$$

$$\frac{6}{5} p = 2q$$

$$3p = 5q$$

$$3p = 5(1 - p)$$

$$3p = 5 - 5p$$

$$8p = 5$$

$$p = \frac{5}{8} \Rightarrow q = 1 - p = 1 - 5/8 = \frac{3}{8}$$

$$\text{P[getting x even numbers]} = P(X = x) = n_{c_x}\, p^x q^{n-x}, x = 0,1,2...,n$$

Therefore the required number of times that in 10,000 sets of 10 throws each we get no even number

$$= 10{,}000 * \text{P[X = 0]} = 10{,}000 * 10_{c_0} (\frac{5}{8})^0 (\frac{3}{8})^{10-0}$$

$$= 0.5499 \text{ approximately}$$

POISSON DISTRIBUTION

Definition

If X is a discrete random variable that assumes only non-negative values such that its probability mass function is given by

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}, \text{ x} = 0,1,2,3,\ldots \quad \text{where} \lambda > 0$$
$$= 0, \text{otherwise}$$

then X is said to follow Poisson distribution with the parameter λ .

Poisson Distribution is a Limiting case of Binomial Distribution

Suppose in a binomial distribution,

1. The number of trails n is indefinitely large, i.e., $n \to \infty$.

2. The probability of success p for each trail is very small, i.e, $p \to 0$.

3. $np(= \lambda)$ is finite and $p = \frac{\lambda}{n}$, $q = 1 - p = 1 - \frac{\lambda}{n}$ where λ is a positive constant.

Proof

Let X be a binomially distributed random variable. Then probability mass function of a binomial distribution is

$$P(X = x) = n_{c_x} p^x q^{n-x}, \text{ x} = 0,1,2,\ldots,\text{n}$$

$$= \frac{n!}{(n-x)!x!} p^x (1-p)^{n-x}$$

$$= \frac{1.2.3\ldots(n-(x+1))(n-x)(n-(x-1)\ldots(n-1)n}{1.2.3\ldots.(n-x)x!} p^x (1-p)^{n-x}$$

We know that mean of binomial distribution =np

Let $np = \lambda \Rightarrow p = \frac{\lambda}{n}$ and $q = 1 - p = 1 - \frac{\lambda}{n}$

$$= \frac{.(n-(x-1))(n-(x-2)\ldots(n-1)n\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^{n-x}}{x!}$$

$$= \frac{n.n(1-1/n)n(1-2/n)...n(1-((x-1)/n))\dfrac{\lambda^x}{n^x}.(1-\dfrac{\lambda}{n})^{n-x}}{x!}$$

$$= \frac{n^x(1-1/n)(1-2/n)...(1-((x-1)/n))\dfrac{\lambda^x}{n^x}.(1-\dfrac{\lambda}{n})^{n-x}}{x!}$$

$$P(X=x) = \frac{(1-1/n)(1-2/n)...(1-((x-1)/n))\lambda^x.(1-\dfrac{\lambda}{n})^{n-x}}{x!}$$

Taking the limit as $n \to \infty$ in the above equation we get

$$P(X=x) = \lambda^x \underset{n\to\infty}{Lt} \frac{(1-\dfrac{\lambda}{n})^{n-x}}{x!} = \frac{e^{-\lambda}\lambda^x}{x!}$$

Mean of the Poisson distribution

Mean = $E(X) = \sum_{x=0}^{\infty} xP(X=x)$

$$= \sum_{x=0}^{\infty} x\frac{e^{-\lambda}\lambda^x}{x!}$$

$$= \lambda e^{-\lambda} \sum_{x=1}^{\infty} \frac{x\lambda^{x-1}}{x!}$$

$$= \lambda e^{-\lambda} \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!}$$

$$\text{Mean} = \lambda e^{-\lambda} e^{\lambda}$$
$$= \lambda$$

Variance of Poisson distribution

$$\text{Var(X)} = E(X^2) - [E(X)]^2$$

$$[E(X)]^2 = \sum_{x=0}^{\infty} x^2 p(x)$$

$$= \sum_{x=0}^{\infty} x^2 \frac{e^{-\lambda}\lambda^x}{x!}$$

$$= \sum_{x=0}^{\infty} (x^2 + x - x)\frac{e^{-\lambda}\lambda^x}{x!}$$

$$= \sum_{x=0}^{\infty} (x(x-1) + x)\frac{e^{-\lambda}\lambda^x}{x!}$$

$$= \sum_{x=0}^{\infty} x(x-1)\frac{e^{-\lambda}\lambda^x}{x!} + \sum_{x=0}^{\infty} x\frac{e^{-\lambda}\lambda^x}{x!}$$

$$= \sum_{x=2}^{\infty} \frac{e^{-\lambda}\lambda^x}{(x-2)!} + \sum_{x=1}^{\infty} \frac{e^{-\lambda}\lambda^x}{(x-1)!}$$

$$= \lambda^2 \sum_{x=2}^{\infty} \frac{e^{-\lambda}\lambda^{x-2}}{(x-2)!} + \lambda \sum_{x=1}^{\infty} \frac{e^{-\lambda}\lambda^{x-1}}{(x-1)!}$$

$$= \lambda^2 e^{-\lambda} \sum_{x=2}^{\infty} \frac{\lambda^{x-2}}{(x-2)!} + \lambda e^{-\lambda} \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!}$$

$$= \lambda^2 e^{-\lambda} e^{\lambda} + \lambda e^{-\lambda} e^{\lambda}$$

$$E(X^2) = \lambda^2 + \lambda$$

$$\text{Var(X)} = E(X^2) - [E(X)]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$$

Therefore variance of the poisson distribution is λ

Moment Generating function of Poisson distribution

Find the moment generating function of the Poisson distribution and hence find the mean and variance

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}, \text{ x} = 0,1,2,3,...$$

The moment generating function of the poisson distribution is

$$M_x(t) = E(e^{tx})$$

$$= \sum_{x=0}^{\infty} e^{tx} p(x)$$

$$= \sum_{x=0}^{\infty} e^{tx} \frac{e^{-\lambda}\lambda^x}{x!}$$

$$= \sum_{x=0}^{\infty} \frac{e^{xt-\lambda}\lambda^x}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{e^{tx}\lambda^x}{x!}$$

$$= e^{-\lambda} \sum_{x=0}^{\infty} \frac{(\lambda e^t)^x}{x!} = e^{-\lambda} e^{\lambda e^t}$$

$$= e^{\lambda(e^t - 1)}$$

To find mean and variance using MGF

$$Mean = E(X) = \mu_1' = \left[\frac{d}{dt}[M_x(t)]\right]_{t=0}$$

$$= \left[\frac{d}{dt}(e^{-\lambda} e^{\lambda e^t})\right]_{t=0}$$

$$= \left[e^{-\lambda} e^{\lambda e^t} \lambda e^t\right]_{t=0}$$

$$= e^{-\lambda} e^{\lambda} \lambda = \lambda$$

$$\text{Var(X)} = E(X^2) - [E(X)]^2$$

$$E(X^2) = \mu_2' = \left[\frac{d^2}{dt^2}[M_x(t)]\right]_{t=0}$$

$$= \left[\frac{d}{dt} [\lambda e^{-\lambda} e^{\lambda e^t} e^t] \right]_{t=0}$$

$$= \left[\lambda e^{-\lambda} [e^{\lambda e^t} e^t + e^t e^{\lambda e^t} \lambda e^t \right]_{t=0}$$

$$= \lambda e^{-\lambda} (e^{\lambda} + \lambda e^{\lambda})$$

$$= \lambda e^{-\lambda} e^{\lambda} (1 + \lambda)$$

$$= \lambda (1 + \lambda) = \lambda + \lambda^2$$

$$E(X^2) = \lambda + \lambda^2$$

$$\text{Var(X)} = E(X^2) - [E(X)]^2 = \lambda + \lambda^2 - \lambda^2 = \lambda$$

For Poisson distribution

Mean=λ

Variance=λ

1.If X is a Poisson variate such that P(X=1)=3/10 and P(X=2)=1/5. Find P(X=0) and P(X=3)

Solution

$$P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!}$$

$$P(X = 1) = \frac{e^{-\lambda} \lambda^1}{1!} = \frac{3}{10} \qquad (1)$$

$$P(X = 2) = \frac{e^{-\lambda} \lambda^2}{2!} = \frac{1}{5} \qquad (2)$$

$$\frac{(2)}{(1)} \quad \Rightarrow \quad \frac{\frac{e^{-\lambda} \lambda^2}{2!}}{\frac{e^{-\lambda} \lambda}{1!}} = \frac{\frac{1}{5}}{\frac{3}{10}}$$

$$\frac{(2)}{(1)} \quad \Rightarrow \qquad \frac{\lambda}{2}=\frac{2}{3} \quad \Rightarrow \lambda=\frac{4}{3}$$

$$P(X=x)=\frac{e^{-\lambda}\lambda^{x}}{x!}=\frac{e^{-\frac{4}{3}}\left(\frac{4}{3}\right)^{x}}{x!}$$

$$P(X=0)=\frac{e^{-\frac{4}{3}}\left(\frac{4}{3}\right)^{0}}{0!}=e^{-\frac{4}{3}}=0.2637$$

$$P(X=3)=\frac{e^{-\frac{4}{3}}\left(\frac{4}{3}\right)^{3}}{3!}=0.1047$$

2. In a certain factory producing razor blades, there is a small chance 1/500 for any blade to be defective. The blades are supplied in packets of 10.Use Poisson distribution to calculate the approximate number of packets containing

(i) no defective blade

(ii) at least 1 defective blade and

(iii) at most 1 defective blade in a consignment of 10,000 packets.

Solution

Given p=1/500 and n=10

Let X be the number of defectives in a packet

λ=np=10/500=1/50=0.02

$$P(X=x)=\frac{e^{-\lambda}\lambda^{x}}{x!}=\frac{e^{-0.02}(0.02)^{x}}{x!}$$

i) No defective blade : P(X=0)

$$=\frac{e^{-0.02}(0.02)^{0}}{0!}=0.9802$$

Therefore the number of pockets containing no defective razor= 10000 * 0.9802

= 9802

ii) At least 1 defective= P(X≥1)

= 1- P(X<1)

= 1-P(X=0)

= 1 – 0.9802 = 0.0198

Therefore the number of packets containing at least one defective= 10000 * 0.0198

= 198

iii) At most 1 defective = P(X≤1)

=P(X=0) + P(X=1)

$$= \frac{e^{-0.02}}{0!} + \frac{e^{-0.02}(0.02)}{1!}$$

$$= 0.0198 + e^{-0.02}(0.02)$$

$$= 0.9997$$

Therefore the number of packets containing at most 1 defective blade = 10000 * 0.9997

= 9997

3. An insurance company has discovered that only about 0.1% of the population is involved in a certain type of accident each year. If its 10000 policy holders were randomly selected from the population, what is the problem that not more than 5 of its clients are involved in such an accident next year?

Solution

Given p= 0.1% = 0.1/100 = 0.001

n= 10000

Mean $\lambda = np = 10000 * 0.001 = 10$

Let X be a random variable of number of clients involved in accident

$$P(X=x)=\frac{e^{-\lambda}\lambda^{x}}{x!}=\frac{e^{-10}(10)^{x}}{x!}$$

$$P(X\le 5)=P(X=0)+P(X=1)+P(X=2)+P(X=3)+P(X=4)+P(X=5)$$

$$P(X\le 5)=\frac{e^{-10}(10)^{0}}{0!}+\frac{e^{-10}(10)^{1}}{1!}+\frac{e^{-10}(10)^{2}}{2!}+\frac{e^{-10}(10)^{3}}{3!}+\frac{e^{-10}(10)^{4}}{4!}+\frac{e^{-10}(10)^{5}}{5!}$$

$$=e^{-10}\left\{1+\frac{10}{1}+\frac{100}{2}+\frac{1000}{6}+\frac{10000}{24}+\frac{100000}{120}\right\}$$

$$=0.0671$$

4. In a given city 4% of all licenced drivers will be involved in at least 1 road accident in any given year. Determine the probability that among 150 licenced drivers ran only chosen in this city

i) only 5 will be involved in atleast 1 accident in any given year and

ii)at most 3 will be involved in atleast 1 accident in any given year.

Solution

$$\lambda=np=100\times\frac{4}{100}=6$$

i) $$P(X=5)=\frac{e^{-6}6^{5}}{5!}=0.1606$$

ii) $$P(X\le 3)=P(X=0)+P(X=1)+P(X=2)+P(X=3)$$

$$=e^{-6}+\frac{e^{-6}6}{1!}+\frac{e^{-6}6^{2}}{2!}+\frac{e^{-6}6^{3}}{31}=0.1512$$

5.In an industrial complex, the average number of fatal accidents per month is one-half. The number of accidents per month is adequately described by a Poisson distribution. What is the probability that 6 months will pass without a fatal accident.

Solution

The average number of fatal accidents per month is $\lambda=1/2$

During 6 months, the average number of fatal accidents would be 1/2+1/2+1/2+1/2+1/2+1/2=3, by additive property of Poisson distribution. λ=3

The probability that 6 months will pass without a fatal accident = P(X=0)

$$P(X=x)=\frac{e^{-\lambda}\lambda^{x}}{x!}$$

$$P(X=0)=\frac{e^{-3}3^{0}}{0!}=0.0498$$

6. A radioactive source emits on the average 2.5 particles per second. Find the probability that 3 or more particles will be emitted in an interval of 4 seconds.

Solution

$\lambda = 2.5/\text{sec}$

In an interval of 4 second average number of particles emitted= 2.5+2.5+2.5+2.5

$= 10$

$$P(3 \text{ or more particles emitted}) = 1 - P[X < 3]$$

$$= 1 - \{P[X=0] + P[X=1] + P[X=2]\}$$

$$= 1 - \left\{\frac{e^{-10}(10)^{0}}{0!} + \frac{e^{-10}(10)^{1}}{1!} + \frac{e^{-10}(10)^{2}}{2!}\right\}$$

$$= 1 - e^{-10}(1 + 10 + \frac{100}{2})$$

$$= 0.9972$$

7. Messages arrive at a switch board in a Poisson manner at an average rate of six per hour. Find the probability for each of the following events

(i) Exactly two messages arrive within one hour

(ii) No message arrives with in one hour

(iii) at least three messages arrive within one hour

Solution

Mean $\lambda = 6$ per hour

$$P(X=x)=\frac{e^{-\lambda}\lambda^{x}}{x!}=\frac{e^{-6}6^{x}}{x!}$$

$$P(X=2)=\frac{e^{-6}6^{2}}{2!}=0.0446$$

$$P(X=0)=\frac{e^{-6}6^{0}}{0!}=0.0025$$

$$P(X\geq 3)=1-P(X<3)=1-[P(X=0)+P(X=1)+P(X=2)]$$

$$=1-e^{-6}(1+6+18)=0.9380$$

8. A car hire firm has 2 cars which it hires out day by day. The number of demands for a car on each day follows a Poisson distribution with mean 1.5. Calculate the proportion of days on which

i) neither car is used

ii) some demand is not fulfilled

Solution

Let X be random variable representing the number of demands for cars:

P(x demands in a day)= $P(X=x)=\frac{e^{-\lambda}\lambda^{x}}{x!}$

Given: $\lambda = 1.5$

Now $P(X=x)=\frac{e^{-1.5}(1.5)^{x}}{x!}$

i) the proportion of days on which neither car is used

$$P(X=0)=\frac{e^{-1.5}1.5^{0}}{0!}=e^{-1.5}=0.2231$$

ii) The proportion of days on which some demand is refused

The demand is refused when x is more than 2

$$P(X>2)=1-[P(X\leq 2)]$$

$$= 1 - [P(X = 0) + P(X = 1) + P(X = 2)]$$

$$= 1 - [\frac{e^{-1.5}(1.5)^0}{0!} + \frac{e^{-1.5}(1.5)^1}{1!} + \frac{e^{-1.5}(1.5)^2}{2!}]$$

$$= 0.19126$$

9. The proofs of a 500 page book contains 500 misprints. Find the probability that there are at least 4 misprints in a randomly chosen page.

Solution

Total number of mistakes= 500

Total number of pages= 500

The average number of mistake per page is 1. $\lambda = 1$

Let X be a random variable of number of mistakes in a page.

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{e^{-1}1^x}{x!}$$

$$\text{P(at least 4 mistakes)} = P(X \geq 4)$$

$$= 1 - P(X < 4)$$

$$= 1 - [P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)]$$

$$= 1 - \left\{\frac{e^{-1}}{0!} + \frac{e^{-1}}{1!} + \frac{e^{-1}}{2!} + \frac{e^{-1}}{3!}\right\}$$

$$= 1 - e^{-1}\left\{1 + 1 + \frac{1}{2} + \frac{1}{6}\right\}$$

$$= 0.0180$$

10. It has been established that the number of defective stereos produced daily at a certain plant is Poisson distributed with mean 4. Over a 2 day span, what is the probability that the number of defective stereos does not exceed 3?

Solution

Let X_1 be the number of defective stereos produced on the first day and X_2 be the number of defective stereos produced on the second day.

Then X_1+X_2 is the number of defective stereos produced on the 2 days

X_1+X_2 follows a poisson distribution with parameter 4+4=8. Thus

$$P(X=x)=\frac{e^{-\lambda}\lambda^{x}}{x!}=\frac{e^{-8}8^{x}}{x!},x=0,1,2..$$

$$P[Number\ \text{of defectives does not exceed}\ 3]=P(X_1+X_2\le 3)$$
$$=P(X_1+X_2=0)+P(X_1+X_2=1)+P(X_1+X_2=2)+P(X_1+X_2=3)$$

$$=\frac{e^{-8}8^{0}}{0!}+\frac{e^{-8}8^{1}}{1!}+\frac{e^{-8}8^{2}}{2!}+\frac{e^{-8}8^{3}}{3!}$$

$$=0.0424$$

11. If the number of telephone calls coming into a telephone exchange between 9 A.M and 10 A.M and between 10 A.M and 11 A.M are independent and follows Poisson distribution with parameters 2 and 6 respectively. What is the probability that more than 5 calls come between 9 A.M and 11 A.M

Solution

Let X be a random variable which denotes the number of telephone calls between 9 am and 10 am with parameter 2 and Y be a random variable which denotes the number of telephone calls between 10 am and 11 am with parameter 6

By additive property X+Y=Z be a random variable having the mean 8 (2+6)

Hence

$$P(Z>5)=1-P(Z\le 5)$$

$$=1-\left[P(Z=0)+P(Z=1)+P(Z=2)+P(Z=3)+P(Z=4)+P(Z=5)\right]$$

$$=1-\sum_{z=0}^{5}\frac{e^{-\lambda}\lambda^{z}}{z!}$$

$$=1-e^{-8}[1+\frac{8}{1!}+\frac{8^2}{2!}+\frac{8^3}{3!}+\frac{8^4}{4!}+\frac{8^5}{5!}]$$

$$= 1 - e^{-8}[1 + 8 + 32 + 85.3333 + 170.669 + 273.0667]$$

$$= 0.80876$$

12. Fit a poisson distribution to the following data and calculate the theoretical frequencies

Deaths	0	1	2	3	4
Frequency	122	60	15	2	1

Solution

X	F	Fx	Theoretical frequencies
0	122	0	121
1	60	60	61
2	15	30	15
3	2	6	3
4	1	4	0

Mean $\bar{x} = \frac{\sum fx}{N} = \frac{100}{200} = 0.5$

Theoretical distribution is given by

$$= N \times P[X = x] = 200 \times \frac{e^{-\lambda}\lambda^{x}}{x!}$$

Hence the theoretical frequencies are given by

$$f(x) = 200 \times \frac{e^{-0.5}(0.5).^{x}}{x!} \qquad (1)$$

Putting x=0,1,2,3,4 in (1) we get

$$f(0) = 200 \times \frac{e^{-0.5}(0.5).^{0}}{0!} = 121$$

$$f(1) = 200 \times \frac{e^{-0.5}(0.5).^{1}}{1!} = 61$$

$$f(2) = 200 \times \frac{e^{-0.5}(0.5)^{2}}{2!} = 15$$

$$f(3) = 200 \times \frac{e^{-0.5}(0.5)^{3}}{3!} = 3$$

$$f(4) = 200 \times \frac{e^{-0.5}(0.5)^{4}}{4!} = 0$$

EXERCISES:

1. The number of typing mistakes that a typist makes on a given page has a Poisson distribution with a mean of 3 mistakes. What is the probability that she makes

i) Exactly 7 mistakes

ii) Fewer than 4 mistakes

iii) No mistakes on a given page

[Answer: 0.0216; 0.0474 ; 0.0498]

2.The number of black floes on a broad bean leaf follows a Poisson distribution with mean 2. A plant inspector however records the number of flies on a leaf only if at least 1 fly is present. What is the probability that he records 1 or 2 flies on a randomly chosen leaf? What is the expected number of flies recorded per leaf?

3. Letters were received in an office an each of 100 days. Assuming the following data to form a random sample from a Poisson distribution, find the expected frequencies, correct to nearest unit.

No. of letters	0	1	2	3	4	5	6	7	8	9	10
Frequencies	1	4	1	2	2	2	8	6	2	0	1

5 2 1 0

NEGATIVE BINOMIAL DISTRIBUTION

The negative Binomial distribution is a reversal of binomial distribution.

The negative Binomial random variable represents the number of failures before the r^{th} success. Here the number of success is fixed.

The negative Binomial variable arises in situations where

i) The experiment consists of a series of independent and identical Bernoulli's trails, each with probability p of success,

ii) The trails are observed until exactly r successes are obtained where r is fixed by the experiment.

iii) The random variable X is the number of failures before the r^{th} success.

Definition

A random variable X is said to assume the negative binomial distribution, if it's probability mass function is given by

$$P[X = x] = x + r - 1_{c_{r-1}}\, p^r q^x, x = 0,1,2,...$$
$$= -r_{c_x}\, p^r (-q)^x, \text{ x} = 0,1,2,...$$

Note

In negative binomial the moments are (r,p)

Moments of Negative Binomial distribution

$$\mu_r{}' = \sum_{x=0}^{\infty} x^r p(x)$$

W.k.t.,

$$\mu_1{}' = E(X) = \sum_{x=0}^{\infty} xp(x)$$

$$= \sum_{x=0}^{\infty} x\left(x + r - 1_{c_{r-1}}\, p^r q^x\right)$$

$$= 0 + rp^r q + \frac{2(r+1)\cdot r}{2!} p^r q^2 + \ldots$$

$$= rp^r q\left[1 + \frac{r+1}{1!} q + \frac{(r+1)(r+2)}{2!} q^2 + \ldots\right]$$

$$= rp^r q\left[\frac{1}{p^{r+1}}\right] = \frac{rq}{p}.$$

$\therefore \mu_1' = \frac{rq}{p}$ i.e., Mean = $\frac{rq}{p}$.

$$\mu_2' = E(X^2) = \sum_{x=0}^{\infty} x^2 p(x)$$

$$= \sum_{x=0}^{\infty} x^2 \, (x+r-1)_{c_{r-1}} p^r q^x$$

$$= \sum_{x=0}^{\infty} [x + x(x-1]x + r - 1_{c_{r-1}} p^r q^x$$

$$= \sum_{x=0}^{\infty} x\left[x + r - 1_{c_{r-1}} p^r q^x\right] + \sum_{x=0}^{\infty} x(x-1)\left[x + r - 1_{c_{r-1}} p^r q^x\right]$$

$$= \frac{rq}{p} + \sum_{x=0}^{\infty} x(x-1)\left[x + r - 1_{c_{r-1}} p^r q^x\right]$$

$$= \frac{rq}{p} + \left[0 + 0 + 2(1)\frac{(r+1)r}{2!} p^r q^2 + 3(2)\frac{(r+2)(r+1)r}{3!} p^r q^3 + 4(3)\frac{(r+3)(r+2)(r+1)r}{4!} p^r q^4 + \ldots\right]$$

$$= \frac{rq}{p} + (r+1)rp^r q^2\left[1 + (r+2)q + \frac{(r+2)(r+3)}{2!} q^2 + \ldots\right]$$

$$= \frac{rq}{p} + (r+1)rp^r q^2 \frac{1}{(1-q)^{r+2}}$$

$$= \frac{rq}{p} + (r+1)rp^r q^2 \frac{1}{p^{r+2}}$$

$$= \frac{rq}{p} + (r+1)r\frac{q^2}{p^2}$$

$$\text{Var(X)} = \text{E(X}^2) - \left[E(X)\right]^2$$

$$= \frac{rq}{p} + (r+1)r\frac{q^2}{p^2} - \left[\frac{rq}{p}\right]^2$$

$$= \frac{rq}{p} + \frac{r^2q^2}{p^2} + \frac{rq^2}{p^2} - \frac{r^2q^2}{p^2}$$

$$= \frac{rq}{p} + \frac{rq^2}{p^2}$$

$$= \frac{rq}{p^2}(p+q)$$

$$= \frac{rq}{p^2}$$ (since p+q=1)

Variance= $E[X^2]-[[E[X]]^2$

Therefore, Variance $= \frac{rq}{p^2}$.

Moment Generating Function of negative binomial distribution

By definition of M.G.F, $$M_X(t) = E(e^{tx}) = \sum_{x=0}^{\infty} e^{tx} p(x)$$

$$= \sum_{x=0}^{\infty} e^{tx} x + r - 1_{c_{r-1}} p^r q^x$$

$$= p^r \sum_{x=0}^{\infty} x + r - 1_{c_{r-1}} \left(qe^t\right)^x$$

$$= p^r\left[1 + rqe^t + \frac{(r+1)r}{2!}(qe^t)^2 + \ldots\right]$$

$= p^r\left(1-qe^t\right)^{-r}$ $\quad (1-x)^{-n} = 1 + n_{c1}x + n_{c2}x^2 + n_{c3}x^3 + ...)$

(

$$\therefore M_X(t) = \frac{p^r}{(1-qe^t)^r} = \left(\frac{p}{1-qe^t}\right)^r .$$

Examples

1.Find the probability that a person tossing 3 coins will get either all heads or all tails for the second time in a fifth toss.

Solution

The sample space is

S={HHH,HHT,HTH,THH,HTT,THT,TTH,TTT}

P[getting all heads or all tails]=P[getting all heads]+P[getting all tails]

$$= \frac{1}{8} + \frac{1}{8} = \frac{2}{8} = \frac{1}{4}$$

p= $\frac{1}{4}$

$$q = 1 - p = \frac{3}{4}$$

Let X be the number of failures before the second success.

$$P(X = x) = x + r - 1_{c_{r-1}}\, p^r q^x$$

r =2, p=1/4. q=3/4

x+r=5 $\Rightarrow x = 5 - 2 = 3$

$$P(X = 3) = 3 + 2 - 1_{c_{2-1}}\, (1/4)^2 (3/4)^3$$

$= 4_{c_1}\, (1/4)^2 (3/4)^3$ =0.1055

2. In a company 5% defective components are produced. What is the probability that atleast 5 components are to be examined in order to get 3 defective ones?

Solution

$p = 5\% = 0.05$

q=0.95

Required probability= $P[X = x + r \geq 5]$

$= P[x + 3 \geq 5]$

$= P[x \geq 2]$

$= 1 - P[x < 2]$

$= 1 - [P[x = 0] + P[x = 1]]$

$= 1 - \{0 + 3 - 1_{c_{3-1}} (0.05)^3 (0.95)^0 + 1 + 3 - 1_{c_{3-1}} (0.05)^3 (0.95)\}$

$= 1 - \{2_{c_2} (0.05)^3 + 3_{c_2} (0.05)^3 (0.95)\}$

$1 - 0.00048 = 0.99952$

3. The probability that a child exposed to certain contagious disease is 0.4. what will be the probability that the tenth child exposed to the disease will be the third to catch it?

Solution

P=0.4, x=10, r=3

q=0.6

$P[X = 10] = P[x + r = 10]$

$= P[x = 10 - r = 10 - 3 = 7]$

$= P[x = 7]$

We know that $P(X = x) = x + r - 1_{c_{r-1}} p^r q^x$

$P[x = 7]$ $= 7 + 3 - 1_{c_{3-1}} (0.4)^3 (0.6)^7$

$= 9_{c_2} (0.4)^3 (0.6)^7$ =0.0645

4. A fair dice is toss on successive independent trails until the second 6 is observed. Find the probability of observing exactly 10 non 6's before the second 6 is toss.

Solution

Success is getting 6 when a dice is toss.

Probability of getting 6= 1/6

P=1/6; q=1-p=1-1/6=5/6

Let X b the number of failures.

$p = \frac{1}{6},\ \mathrm{q} = \frac{5}{6}, X = 10,\ r = 2$

$P[x = 10] = 10 + 2 - 1_{c_{2-1}} (1/6)^2 (5/6)^{10}$

$= 11_{c_1} (1/6)^2 (5/6)^{10}$

$= 11(0.0045) = 0.04935$

5. If the probability of a male child is $1/2$, then find the probability that in a family the 6th child is the third female child?

Solution

Here we have to find the probability that the sixth child is the third female child.

We have to apply negative Binomial distribution.

$P(X = x) = x + r - 1_{\epsilon_{r-1}} p^r q^x$

P=Probability of female child =1/2

q=1-p=1/2

r=3

x+r=6

x=6-3=3

$$P[x=3]=3+3-1_{c_{3-1}}(1/2)^3(1/2)^3$$

$$=5_{c_2}(1/2)^3(1/2)^3=5_{c_2}(1/2)^3=0.15625$$

ERLANG DISTRIBUTION OR GENERAL GAMMA DISTRIBUTION

Definition

A continuous random variable X is said to follow an erlang distribution or General Gamma distribution with parameters $\lambda>0$ and $k>0$ if its p.d.f is given by

$$f(x)=\frac{\lambda^k x^{k-1}e^{-\lambda x}}{\Gamma k},\ \text{x}\geq 0$$
$$=0,\text{otherwise}$$

Note

1. When $\lambda=1$, the Erlang distribution is called Gamma distributionr simple Gamma distribution with

Parameter k whose density function is

$$f(x)=\frac{1}{\Gamma k}x^{k-1}e^{-x},\ \text{x}\geq 0,\text{k}>0$$

2. when k =1, the Erlang distribution leads to the exponential distribution.

Moments of Gamma Distribution

$$\mu_r{}' = E(x^r)=\int_{-\infty}^{\infty}x^r f(x)dx$$

$$=\int_0^{\infty}x^r\frac{\lambda^k x^{k-1}e^{-\lambda x}}{\Gamma k}dx$$

$$=\frac{\lambda^k}{\Gamma k}\int_0^{\infty}x^{r+k-1}e^{-\lambda x}dx$$

$$= \frac{\lambda^k}{\Gamma k} \frac{\Gamma k + r}{\lambda^{k+r}}$$

$$\mu_r{}' = \frac{\Gamma k + r}{\lambda^r \Gamma k}$$

Putting r=1

$$\mu_1{}' = \frac{\Gamma k + 1}{\lambda \Gamma k} = \frac{k \Gamma k}{\lambda \Gamma k} = \frac{k}{\lambda}.$$

Mean= E(X)= $\frac{k}{\lambda}$

$$\mu_2{}' = \frac{\Gamma k + 2}{\lambda^2 \Gamma k} = \frac{k+1 \Gamma k + 1}{\lambda^2 \Gamma k} = \frac{(k+1) k \Gamma k}{\lambda^2 \Gamma k}$$

$$= \frac{k(k+1)}{\lambda^2}$$

Var(X)= $\mu_{2'} - (\mu_1{}')^2$

$$= \frac{k(k+1)}{\lambda^2} - \left(\frac{k}{\lambda}\right)^2$$

$$\frac{k^2 + k = k^2}{\lambda^2} = \frac{k}{\lambda^2}$$

Var(X)= $\frac{k}{\lambda^2}$

GAMMA DISTRIBUTION

Definition

A continuous random variable X is said to follow the Gamma distribution with parameter λ. If the probability density function is given by

$$f(x) = \frac{e^{-x}x^{\lambda-1}}{\Gamma\lambda}, \lambda > 0, 0 < x < \infty$$
$$= 0, \text{otherwise}$$

Moment Generating function of Gamma distribution

$$M_X(t) = E(e^{tx}) = \int_{-\infty}^{\infty} e^{tx} f(x)dx$$

$$= \int_0^{\infty} e^{tx} \frac{e^{-x}x^{\lambda-1}}{\Gamma\lambda} dx$$

$$= \frac{1}{\Gamma\lambda} \int_0^{\infty} e^{(t-1)x} x^{\lambda-1} dx$$

$$= \frac{1}{\Gamma\lambda} \frac{\Gamma\lambda}{(1-t)^{\lambda}}$$

$$= \frac{1}{(1-t)^{\lambda}}$$

$$M_X(t) = (1-t)^{-\lambda}$$

Mean and Variance of Gamma Distribution

$$\mu_1' = E(X) = \left[\frac{d}{dt} M_X(t)\right]_{t=0}$$

$$= \left[\frac{d}{dt}(1-t)^{-\lambda}\right]_{t=0}$$

$$= \left[-\lambda(1-t)^{-\lambda-1}(-1)\right]_{t=0}$$

Mean= λ .

$$\mu_2' = E(X^2) = \left[\frac{d^2}{dt^2} M_X(t)\right]_{t=0}$$

$$= \frac{d}{dt}\left[\lambda(1-t)^{-(\lambda+1)}\right]_{t=0}$$

$$= \left[\lambda(\lambda+1)(1-t)^{-(\lambda+2)}\right]_{t=0}$$

$$\mu_2' = \lambda(\lambda+1)$$

$$\text{Var(X)} = \text{E(X}^2) - \left[E(X)\right]^2$$

$$= \lambda(\lambda+1) - \lambda^2 = \lambda$$

$$\text{Var(X)} = \lambda$$

Examples

1. The daily consumption of milk in excess of 20,000 litres is approximately distributed as Gamma variable with parameter k=2, $\lambda = \frac{1}{10{,}000}$. If the city has a daily stock of 30,000 litres on a given day, find the probability that the stock is insufficient.

Solution

Let X be the random variable for daily consumption of milk in the city.

Let Y=X-20,000. Then Y is a Gamma variable.

$$f(Y) = \frac{\lambda^k y^{k-1} e^{-\lambda y}}{\Gamma k} \ , \ \text{y} \geq 0$$

When k =2, $\lambda = \frac{1}{10{,}000}$

$$f(Y) = \frac{\left(\frac{1}{10{,}000}\right)^2 ye^{-\frac{1}{10{,}000}y}}{\Gamma 2}$$

$$P(X > 30{,}000) = P(Y > 10{,}000)$$

$$= \int_{10{,}000}^{\infty} \frac{y}{(10{,}000)^2} e^{-y/10{,}000} dy$$

Put $\frac{y}{10{,}000} = t$, $\frac{dy}{10{,}000} = dt$

Therefore

$$P(X > 30{,}000) = \frac{1}{10{,}000}\int_{1}^{\infty} te^{-t} 10{,}000\, dt$$

$$= \int_{1}^{\infty} te^{-t}\, dt$$

$$-te^{-t} - e^{-t}\Big]_{1}^{\infty}$$

$$-\left[-1e^{-1} - e^{-1}\right] = 2e^{-1}$$

2. The consumer demand for electricity in a certain locality per month is known to follow general Gamma distribution. If the average demand is a kilowatt and the most likely demand is b kilowatt(b<a) What is the variance of the demand.

Solution

Let X be the random variable denoting consumer demand of electricity,

$$f(x) = \frac{\lambda^k}{\Gamma k} x^{k-1} e^{-\lambda x},\ x > 0$$.

The most likely demand is the mode of X. i.e., the value of x for which f(x) is maximum.

i.e., $f'(x) = \frac{\lambda^k}{\Gamma k}\left[(k-1)x^{k-2}e^{-\lambda x} + x^{k-1}(-\lambda e^{=\lambda x})\right]$

$$= \frac{e^{-\lambda}\lambda^k x^{k-2}}{\Gamma k}\left[(k-1) - \lambda x\right]$$

$f'(x) = 0$ when x = 0, $x = \frac{k-1}{\lambda}$

$$f''(x) = \frac{\lambda^k}{\Gamma k}\left[-\lambda x^{k-2}e^{-\lambda x} + \left[(k-1) - \lambda x\right]\frac{d}{dx}\left(x^{k-2}e^{-\lambda x}\right)\right]$$

$$f''(x) = \frac{\lambda^k}{\Gamma k}\left[-\lambda x^{k-2}e^{-\lambda x} + [(k-1)-\lambda x]\frac{d}{dx}\left(x^{k-2}e^{-\lambda x}\right)\right] < 0 \text{ when } x - \frac{k-1}{\lambda}.$$

i.e., $\frac{k-1}{\lambda} = b$

Now $E(X) = \frac{k}{\lambda} = a$ [since average demand =a]

$$a - b = \frac{1}{\lambda}$$

Therefore $\text{Var(X)} = \frac{\text{k}}{\lambda^2} = \left(\frac{k}{\lambda}\right)\left(\frac{1}{\lambda}\right) = a(a-b)$.

3. In a city, the daily consumption of electric power is million kilowatts hours in a random variable with general gamma distribution or Erlang distribution with parameter $\lambda = 1/2$ and k=3. If the power plant of this city has a daily capacity of 12 million kilowatts hours. What is the probability that this power supply will be inadequate on any given day.

Solution

Let X be the random variable that denotes the daily consumption of power. Then

$$f(x) = \frac{\left(\frac{1}{2}\right)^3 x^{3-1}e^{-\frac{1}{2}x}}{\Gamma 3}, \text{ x} \geq 0$$

$$f(x) = \frac{\left(\frac{1}{2}\right)^3 x^{21}e^{-\frac{1}{2}x}}{\Gamma 3}.$$

The power supply will be inadequate if the consumption goes beyond 12 million kilowatt hour.

$$P[X > 12] = \int_{12}^{\infty} f(x)dx$$

$$= \int_{12}^{\infty} \frac{1}{8}x^2 \frac{e^{-\frac{x}{2}}}{\Gamma 3}dx$$

$$= \frac{1}{8}.\frac{1}{\Gamma 3}\int x^2 e^{-x/2} dx$$

$$= \frac{1}{8.2}\left[\frac{x^2 e^{-x/2}}{-1/2} - \frac{2xe^{-x/}\ 2}{(1/2)^2} + \frac{2e^{-x/2}}{-(1/2)^3}\right]_{12}^{\infty}$$

$$= \frac{1}{16}\left[-2x^2 e^{-x/2} - 8xe^{-x/2} - 16e^{-x/2}\right]_{12}^{\infty}$$

$$= \frac{1}{16}(-)\left[-2(12)^2 e^{-6} - 8(12)e^{-6} - 16e^{-6}\right]$$

$$= \frac{1}{16}e^{-16}\left[288 + 96 + 16\right] = 25e^{-6} = 0.0625$$

4. The daily consumption of bread in a hostel in excess of 2000 leaves is approximately gamma distributed with parameters k =2 and $\lambda = \frac{1}{1000}$. The hostel has a daily stock of 3000 leaves. What is the probability that the stock is insufficient on a day.

Solution

Let X be the number of leaves consumed daily .

Let $Y = X - 2000$.

Then Y follows gamma distribution with k =2 and $\lambda = \frac{1}{1000}$.

$$f(Y) = \frac{\lambda^k y^{k-1} e^{-\lambda y}}{\Gamma 2}\ ,\ \text{y} \geq 0$$

$$= \left(\frac{1}{1000}\right)^2 \frac{y^{2-1} e^{-\frac{1}{1000}y}}{\Gamma 2}$$

$$= \left(\frac{1}{1000}\right)^2 ye^{-\frac{1}{1000}y} \quad ,\text{y>0}$$

P[Stock is insufficient]= $P(X > 3000)$

$= P(Y + 2000 > 3000)$

$= P(Y > 1000)$

$$= \int_{1000}^{\infty} \frac{1}{1000^2} y e^{-y/1000} dy$$

$= 2e^{-1} = 0.7358$

EXPONENTIAL DISTRIBUTION

Definition

A continuous random variable X is designed in $(0, \infty)$ is said to follow an exponential distribution if the probability density function is

$$f(x) = \lambda e^{-\lambda x}, \lambda > 0$$
$$x \geq 0$$

Note

In exponential distribution, λ is the parameter

Cumulative distribution function

$$F(x) = \int_0^x \lambda e^{-\lambda x} dx \text{ if } x > 0$$
$$0 \text{ otherwise}$$

$$= 1 - e^{-\lambda x}, \text{ if } x > 0$$
$$0 \text{ otherwise}$$

Moments, mean and variance

The rth moment is given by

$$\mu_r' = E[X^r] = \int_0^{\infty} x^r \lambda e^{-\lambda x} dx$$

$$= \lambda \int_0^{\infty} e^{-\lambda x} x^{r+1-1} dx$$

$$= \frac{\lambda \Gamma(r+1)}{\lambda^{r+1}} = \frac{\Gamma r+1}{\lambda^r} = \frac{r!}{\lambda^r}$$

$$\mu_r{}' = \frac{r!}{\lambda^r}$$

Now, $\mu_1{}' = \frac{1!}{\lambda^1} = \frac{1}{\lambda}$

$$\mu_2{}' = E(X^2) = \frac{2!}{\lambda^2} = \frac{2}{\lambda^2}$$

$$Var(X) = E(X^2) - [E(X)]^2 = \frac{2}{\lambda^2} - (\frac{1}{\lambda})^2$$

$$= \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

$$Var(X) = \frac{1}{\lambda^2}$$

Moment Generating Function, Mean and Variance

$$M_X(t) = E(e^{tx}) = \int_0^{\infty} e^{tx} f(x) dx$$

$$= \int_0^{\alpha} e^{tx} \lambda e^{-\lambda x} dx$$

$$= \lambda \int_0^{\alpha} e^{-(\lambda - t)x} dx$$

$$= \lambda \left[\frac{e^{-e(\lambda - t)x}}{-(\lambda - t)} \right]_0^0 = \frac{\lambda}{\lambda - t}$$

$$\frac{d}{dt}M_x(t) = \lambda\left[\frac{1}{(\lambda - t)^2}\right]$$

$$\therefore \mu_1' = \left[\frac{d}{dt}M_X(t)\right]_{t=0} = \frac{\lambda}{\lambda^2} = \frac{1}{\lambda}$$

$$\frac{d^2}{dt^2}M_x(t) = \frac{2\lambda}{(\lambda - t)^3}$$

$$\mu_2' = \left[\frac{d^2}{dt^2}M_x(t)\right]_{t=0} = \frac{2\lambda}{\lambda^3} = \frac{2}{\lambda^2}$$

$$\text{Mean} = \mu_1' = 1/\lambda$$

$$\text{Variance} = \mu_2' - \mu_1' = \frac{2}{\lambda^2} - \left(1/\lambda\right)^2 = \frac{1}{\lambda^2}$$

Examples

1. Suppose that during a rainy season in a tropical island, the length of the shower has an exponential distribution with average 2 minutes. Find the problem that the shower will be there for more than three minutes. If the shower has already lasted for 2 minutes, what is the problem that it will last for at least one more minute.

Solution

Let x be random variable representing the length of the shower in minute.

Given that X follows exponential distribution

The average length = 2.

Therefore the parameter $\lambda = 2$ (since mean = $1/\lambda$)

Hence the p.d.f is

$$f(x) = \lambda e^{-\lambda x} = \frac{1}{2}e^{-1/2x}, \text{ x} \geq 0$$

$$= 0, \text{otherwise}$$

(i) To find the probability of shower lasting more than 3 minutes.

$$= P(x > 3) = \int_3^{\infty} \frac{1}{2} e^{-x/2} dx$$.

$$= \frac{1}{2}\left[\frac{e^{-x/2}}{-1/2}\right]_3^{\infty}$$

$$= \frac{1}{2}\left[\frac{e^{-3/2}}{-1/2}\right] = 0.2231$$.

ii) P[The shower will last at least one more minutes given that it has lasted 2 minutes]= $P[X > 3 / X > 2]$

=P[X>1]= $\int_1^2 \frac{1}{2} e^{-x/2} dx$

$$= \frac{1}{2}\frac{e^{-x/2}}{-1/2}\Big]_1^{\infty}$$

2. The daily consumption of milk in a city in excess of 20,000 litres is approximately exponentially distributed. The average excess in consumption of milk is 3000 litres. The city has a daily stock of 35,000 litres. What is the probability thst of two days selected at random, the stock is insufficient for both the days.

Solution

Let X be the random variable of daily consumption of milk in excess of 20,000 litres.

X is exponentially distributed with mean 3000 litres.

$\therefore \lambda = 1/3000$

$f(x) = \frac{1}{3000} e^{-x/3000}$, x ≥ 0 .

Let Y denote the daily consumption. Then x=Y-20000

P[the stock is insufficient on any day)=P(the consumption exceeds 35,000 litres)

i.e., P(Y>35,000)=P(X+20000>35,000)

=P(X>15,000)

$$= \int_{15,000}^{\infty} \lambda e^{=\lambda x} dx$$

$$= \int_{15,000}^{\infty} \frac{1}{3000} e^{=x/3000} dx$$

$$= \frac{1}{3000}\left[e^{-\infty} - \frac{e^{-15,000/3000}}{-1/3000} \right]$$

$$= e^{=5}$$

Therefore P[stock is insufficient on both days] $= e^{-5} \cdot e^{-5} = e^{-10}$

3. The mileage which car owners get with a certain type of radial tyre is a random variable having an exponential with mean 40,000 k.m. Find the probability that atleast one of these tyres will last

i) atleasr 20,000k.m

ii) atmost 30,000k.m

Solution

Given mean = 40,000 k.m

$$\lambda = \frac{1}{\text{mean}} = \frac{1}{40,000}$$

$$f(x) = \lambda e^{-\lambda x}, x \geq 0$$

$$= \frac{1}{20000} e^{-\frac{1}{40,000}x}, x \geq 0$$

i) $P[X \geq 20,000] = \int_{20,000}^{\infty} f(x)dx$

$$= \int_{20,000}^{\infty} \frac{1}{40,000} e^{-\frac{1}{40,000}x} dx$$

$$= \frac{1}{40,000}\left[\frac{e^{-\frac{1}{40,000}x}}{-\frac{1}{40,000}}\right]_{20,000}^{\infty}$$

$$= \frac{1}{40,000}\left(-40000(0 - e^{-1/2}\right) = e^{-1/2}$$ =0.6065.

ii) $P[X \le 30,000] = \int_{0}^{30,000} \frac{1}{40,000} e^{-\frac{1}{40,000}x} dx$

$$= \frac{1}{40,000}\left[\frac{e^{-\frac{1}{40,000}x}}{-\frac{1}{40,000}}\right]_{0}^{30,000}$$

$$= \frac{1}{40,000}\left(-40000(e^{-\frac{3}{4}} - 1\right) = 1 - e^{-\frac{3}{4}}$$

=0.5276.

4. The time (in hours) required to repair a machine is exponentially distributed with parameter $\lambda = \frac{1}{2}$. What is the probability that the repair time exceeds 2 hours? What is the conditional probability that the repair time take takes atleast 10 hours given that its duration exceeds 9 hours?

Solution

Let X be a random variable of time to repair the machine,

Given X is exponentially distributed with $\lambda = \frac{1}{2}$.

Therefore $f(x) = \lambda e^{-\lambda x} = \frac{1}{2} e^{-x/2}$, x > 0 .

To find $P[X > 2]$

$$P[X > 2] = \int_2^\infty f(x)dx$$

$$= \int_2^\infty \frac{1}{2} e^{-x/2} dx$$

$$= \frac{1}{2}\left[\frac{e^{-x/2}}{-1/2}\right]_2^\infty$$

$$= -\left[e^{-\infty} - e^{-2/2}\right] = e^{-1} = 0.3679.$$

To find $P[X > 10 / X > 9]$

$P[X > 10 / X > 9] = P[X > 1]$ (By Memory less property)

$$= \int_1^\infty \frac{1}{2} e^{-x/2} dx$$

NORMAL DISTRIBUTION

Definition

A normal distribution is a continuous distribution given by $y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ where X is a continuous normal variate distributed with density function

$f(\lambda) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ with mean μ and standard deviation σ.

Deviation of the distribution

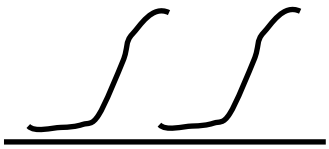
When mean has been taken at the origin but if however another point is taken as the origin such that the excess of the mean over the arbitrary origin is m then

$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2}$ is the standard form of the normal curve with origin at (m,0).

Area under the normal curve is unity.

Characteristics of the Normal Distribution

The diagram of a normal distribution is given below. It is called normal curve.



Properties of the Normal Distribution

1. The normal distribution is a symmetrical distribution and the graph of the normal distribution is bell shaped.

2. The curve has a single peak point (i.e.,) the distribution is unimodal

3. The mean of the normal distribution lies at the centre of normal curve.

4.Because of the symmetry of the normal curve, the median and mode are also at the centre of the normal curve. Hence in a normal distribution the mean, median and mode coincide.

5. The tails of the normal distribution extend indefinitely and never touch the horizontal axes. That is we say that the normal curve approaches approximately from either side of its horizontal axes.

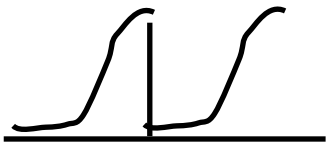
6. The normal distribution is a two parameter probability distribution. The parameters mean and standard deviation (μ,σ) completely determine the distribution.

7. Area property:

In a normal distribution about 67% of the observations will lie between mean $\pm$ S.D i.e., ($\mu \pm \sigma$). About 95% of the observations lie between mean $\pm$ 2S.D (i.e., $\mu \pm 2\sigma$). About 99% of the observation will lie between mean $\pm$ 3S.D i.e.,($\mu \pm 3\sigma$) .

Standard Normal Probability Distribution

If X is a normally distributed random variable, μ and σ are respectively its mean and standard deviation, then $Z = \frac{X-\mu}{\sigma}$ is called standard normal random variable.



Normal table

Special table called table of areas under normal curve is available to determine probabilities that the random variable lies in a given range of values of the variables. Using the table, we can determine the probability for X, taking a value less than x (X<x) and also for a given probability we determine the value x such that X < x

Additive property of Normal Distribution:

If $X_1, X_2, \ldots, X_n$ are independent normal variates with parameters $(m_1,\sigma_1),(m_2,\sigma_2)$

$\ldots,(m_n,\sigma_n)$ respectively then $X_1+X_2+\ldots..+X_n$ is also a normal variate with parameter (m, σ)

Where $m=m_1+m_2+\ldots.+m_n$ and $\sigma^2=\sigma_1^2+\sigma_2^2+\ldots\ldots..+\sigma_n^2$.

Examples

1) X is a normal variate with mean 30 and standard deviation 5. Find the probability that

i) $26 \le X \le 40$; ii) $X \ge 45$ iii) $|X-30|>5$.

Solution

Given $\mu = 30; \sigma = 5$

$$z = \frac{X-\mu}{\sigma}$$

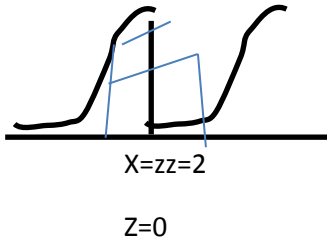
i) when X= 26, $z = \frac{26-30}{5} = -0.8$

when X=40, $z = \frac{40-30}{5} = 2$

$\therefore P[26 \le X \le 40] = P[-0.8 \le z \le 2]$

$= P[-0.8 \le z \le 0] + P[0 \le z \le 2]$

$= P[0 \le z \le 0.8] + P[0 \le z \le 2]$

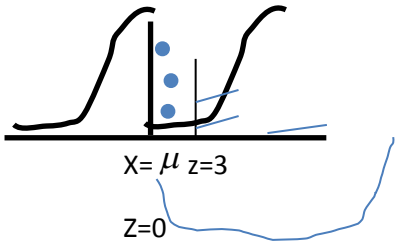


-0.8

=0.2881+0.4772

=0.7653.

ii) when X=45, $z = \dfrac{45-30}{5} = 3$



$\therefore P(X \ge 45) = P(z \ge 3)$

$= P(0 \le Z \le \infty) - P(0 \le Z \le 3))$

$=0.5- P(0 \le z \le 3)$

=0.5-0.4987=0.0013.

iii) To find $P(|X-30| > 5)$

$P(|X-30| \le 5) = P(25 \le X \le 35)$

When X=25, $z = \dfrac{25-30}{5} = -1$

When X=35, $z = \dfrac{35-30}{5} = 1$

$P(|X-30| \le 5) = P(25 \le X \le 35) = P(-1 \le z \le 1) = P(-1 \le z \le 0) + P(0 \le z \le 1) = P(0 \le z \le 1) + P(0 \le z \le 1)$

$= 2P(0 \le z \le 1)$

=2(0.3413)

=0.6826.

$\therefore P(|X-30| > 5) = 1 - P(|X-30| \le 5)$

=1-0.6826

=0.3174

2. A normal distribution has mean $\mu = 20$ and standard deviation $\sigma = 10$ .Find $P(15 \le X \le 40)$.

Solution

Given $\mu = 20$ and $\sigma = 10$

We know that $z = \frac{X - \mu}{\sigma}$.

When $X = 15$, $z = \frac{15 - 20}{10} = -0.5$ and

When =40, $z = \frac{40 - 20}{10} = 2$

$$P(-0.5 \le z \le 2) = P(0 \le z \le 2) + P(0 \le z \le 0.5)$$

=0.4772+0.1915

=0.6687.

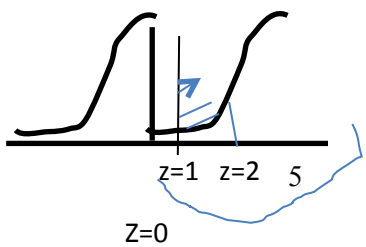
3. The average seasonal rainfall in a place is 16 inches with a standard deviation of 4 inches. What is the probability that in a year the rainfall in that place will be between 20 and 24 inches?

Solution

$$z = \frac{X - \mu}{\sigma}$$

xWhen X=20, $z = \frac{20 - 16}{4} = 1$

When X =24, $z = \frac{24 - 16}{4} = 2$



$$\therefore P(20 < X < 24) = P(1 < z < 2)$$

$$= P(0 < z < 2) - P(0 < z < 1)$$

=0.4772-0.3413

=0.1359.

Note

E(aX+bY)=aE(X)+bE(Y)

Var(aX+bY)=a^2V(X)+b^2V(Y)

Var(a)=0

E(a)=a

4. X is a normal variate with mean 1 and variance 4. Y is another normal variate independent of X with mean 2 and variance 3. What is the distribution of X+2Y?

Solution

Since X and Y are independent normal variates, X+2Y will also be a normal variate by the additive property and

Mean of X+2Y= E(X+2Y)=E(X)+2E(Y) (since E(AX+BY)=AE(X)+BE(Y))

=1+2(2)=5

Variance of X+2Y=V(X+2Y)=V(X)+2^2V(Y) (since Var(AX+BY)=A^2V(X)+B^2V(Y))

$1^2(4)+2^2(3)=16$.

∴ X+2Y will follow normal with mean 5 and variance 16.

5. The saving bank account of a customer showed an average balance of Rs.150 and a standard deviation of Rs.50. Assuming that the account balances are normally distributed.

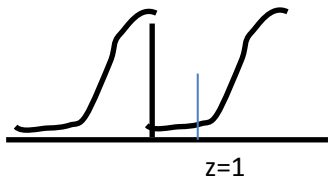
1. What percentage of account is over Rs. 200? P(X>200)
2. What percentage of account is between Rs.120 and Rs.170? P(120<X,170)
3. What percentage of account is less than Rs.75? P(X,75)

Solution

1) To find $P(X \geq 200)$

We know that $z = \frac{X - \mu}{\sigma}$

When X=200, $z = \frac{200 - 150}{50} = 1$



The area from z=0 to infinity is 0.5

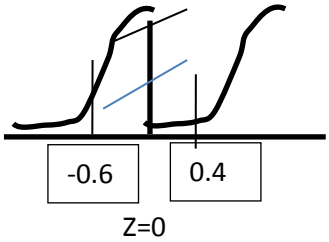
From this subtract area from z=0 to z=1 (this value get it from table)

$P(X \geq 200) = P(z \geq 1) = 0.5 - P(0 < z < 1)$

=0.5-0.3413

=0.1587.

$\therefore$ Percentage of account is over Rs. 200 is 15.87%.



2. To find $P(120 < X < 170)$

When X= 120, $z = \frac{120-150}{50} = -0.6$

When X=170, $z = \frac{170-150}{50} = 0.4$

$\therefore P(120 < X < 170) = P(-0.6 < z < 0.4)$

$= P(0 < z < 0.6) + P(0 < z < 0.4)$

=0.2257+0.1554=0.3811

Therefore, percentage of account between Rs.120 and Rs.170 is 0.3811(100)=38.11.

3. To find $P(X < 75)$

When X=75, $z = \frac{75-150}{50} = -1.5$

$\therefore P(X < 75) = P(z < -1.5)$

$= 0.5 - P(0 < z < 1.5)$

=0.5-0.4322=0.0668.

Therefore, percentage of account is less than Rs.75 is 6.68%

6. The mean yield for one-acre plot is 662 kilos with standard deviation 32 kilos. Assuming normal distribution, how many one-acre plots in a patch of 1000 plots would you expect to have yield over 700 kilos below 650 kilos.

Solution

Given $\mu = 662, \sigma = 32$

$$z = \frac{X-\mu}{\sigma} = \frac{X-662}{32}$$

When X=700, $z = \frac{700-662}{32}$ =1.19

When X=650, $z = \frac{650-662}{32} = -0.375 = -0.38$

$$P[X > 700] = P(z > 1{,}19)$$

$$= 0.5 - P(0 \le z < 0.38) = 0.352$$

Therefore, the number of plots have yield below 650 kilos=352.

Exercises

An electrical firm manufactures light bulbs that have a life, before burnout, that is normally distributed with mean equal to 800 hours and a standard deviation of 40 hours. Find

i) the probability that a bulb burns more than 834 hours.

ii) the probability that bulb burns between 778 and 834 hours.

7. In a distribution exactly normal 7% of the items are under 35 and 89% are under 63. What are the mean and standard deviation of the distribution?

Solution

We know that $z = \frac{X-\mu}{\sigma}$

Let $z=z_1$ when X=35 and $z=z_2$ when X=63

$$P(0 < z < z_1) = 7\% = 0.07$$

$$P(z_1 < z < 0) = 43\% = 0.43$$

From tables, z_1=-1.48

i.e., $\frac{35-\mu}{\sigma} = -1.48$

$$35 - \mu = -1.48\sigma \qquad (1)$$

$P[z > z_1] = 39\% = 0.39$

$P(0 < z < z_2) = 0.39$

From tables, $z_2 = 1.2$

i.e., $\frac{63-\mu}{\sigma} = 1.2$

$63 - \mu = 1.2\sigma$ (2)

(2)-(1) $\Rightarrow 28 = 2.68\sigma$

$\sigma = \frac{28}{2.68} = 10.45$

(2) $\Rightarrow 63 - \mu = (1.2)(10.45)$

$63 - \mu = 12.54$

$\mu = 50.46$

$\therefore$ Mean = 50.5 and standard deviation = 10.5.

8. In a normal distribution, 31% of the items are under 45 and 8% are over 64. Find the mean and variance of the distribution.

Solution

We know that $z = \frac{X-\mu}{\sigma}$.

Let $z=z_1$ when X=45 and $z=z_2$ when X=64

$P(0 < z < z_1) = 31\% = 0.31$

$P(z_1 < z < 0) = 0.19$

From tables, $z_1=-0.49$

i.e., $\frac{45-\mu}{\sigma} = -0.49$

$45 - \mu = -0.49\sigma$ (1)

$P[z > z_1] = 0.8$ or $P[0 < z < z_2) = 0.42$

From tables, $z_2 = 1.40$

i.e., $\frac{64-\mu}{\sigma} = 1.40$

$64 - \mu = 1.40\sigma$ (2)

(1)-(2) $\Rightarrow$ -19=-1.89σ

$\sigma = \frac{-19}{-1.89} = 10$

From (1) $\mu = 45 + (0.49)(10)$

$\mu = 45 + 4.9 = 49.9$

$\mu = 50$

Therefore mean =50 and standard deviation =10

9. Suppose the heights of men of a certain country are normally distributed with average 68 inches and standard deviation 2.5. Find the percentage of men who are

i) between a=66 and b= 71 inches in height.

ii) approximately 6 feet tall

Solution

Given $\mu = 68, \sigma = 2.5$

We know that $z = \frac{X-\mu}{\sigma}$

When X=66, $z = \frac{66-68}{2.5} = -0.8$

When X=71, $z = \frac{71-68}{2.5} = 1.20$

$P[66 \le X \le 71] = P[-0.8 < z < 1.20]$

$= P[-0.8 < z < 0] + P[0 \le z \le 1.20]$

=0.2881+0.3849=0.6730

Approximately 67.3% men are between 66 and 71 inches in height.

ii) Assuming heights are round of the nearest inch.

Then we get a=71.5, b = 72.5

$$P[71.5 \le X \le 72.5] = P\left[\frac{71.5-68}{2.5} \le \frac{X-68}{2.5} \le \frac{72.5-68}{2.5}\right]$$

$= P[1.4 \le z \le 1.8]$

$= P[0 \le z \le 1.8] - P[0 \le z \le 1.4]$=0.0449

UNIT-III

BASIC STATISTICS

Introduction

The word statistics to indicate the numerical data in any field of enquiry and the term statistical method to denote ‘ the technique of studying and analyzing the data.

Variables

Any character which can vary in magnitude or quality is called a variate.

Thus the height, weight, age, intelligence, number of children per family, the number of students in a class etc are variates.

Variates are of two types

i) Discrete, ii) continuous

i) Discrete: Discrete variates are the values obtained by counting

Example: The number of children in a family and the number of students in a class.

They will be in whole number only.

ii) **Continuous Variate**: Continuous variates correspond to variables which are measured theoretically to any degree of accuracy.

Example: Measurement of height, weight, temperature etc.

Frequency tables:

Suppose we know the marks obtained by 140 candidates in an examination in a certain subject. Let the marks be given by the following table

86	35	69	12	55	53	41	10	35	58	71	45	50	30
59	56	37	29	29	51	47	46	82	36	52	59	32	54
16	65	42	27	53	39	40	62	54	53	38	69	66	50
74	26	44	21	53	32	41	54	32	81	58	45	48	30
57	37	43	77	34	21	61	54	46	33	62	52	75	89
66	75	60	47	85	37	49	61	93	50	51	8	45	49
20	70	47	41	49	16	60	63	39	58	23	40	44	68
52	28	52	51	31	83	36	80	70	43	52	35	18	60
27	60	31	44	78	48	55	38	25	59	22	72	61	48
41	98	67	42	42	33	11	64	72	46	37	76	65	43

(Table 1)

Frequency table of marks of 140 candidates

Class limits	Tally Markss	Frequency
0-9	I	1
10-19	𝍸 I	6
20-29	𝍸 𝍸 II	12
30-39	𝍸 𝍸 𝍸 𝍸 III	23
40-49		30
50-59		29
60-69		19
70-79	𝍸 𝍸 I	11
80-89		7
90-99		2
		140

(Table 2)

We have arranged the raw data into classes of appropriate size, showing the corresponding frequency of variates in each class.

When any set of data is symmetrically arranged in this way, it is called a frequency distribution.

In the above the pairs of numbers written in the column of classes are called lower and upper class limits. Some times called open class limits or class boundaries.

The difference between the heighest mark and lowest mark is called the range.

The range is divided into classes of appropriate size. The size of the class is called the class-interval. Usually the classes are of equal size.

The mid value of such a class is called the class-mak, mid-value or the central value.

The width of a class interval is therefore the common difference beween the consecutive class marks. It is also the difference between the lower (or upper) limits of two successive classes.

In the above, the class interval is 10 and the successive class marks are 4.5, 14.5 etc

Cumulative Frequency:

If f1, f2, f3, … are the frequencies of the successive classes, then f1, f1+f2, f1+f2+f3, etc are the cumulative frequencies.

Thus the cumulative frequencies for the previous table is 1, 7, 19, 42, …

The cumulative frequency 42 gives the number of students obtaining 39 or lower marks.

Relative Frequency:

In some problems we may require the relative frequency instead of the actual or absolute frequency.

The relative frequency of any class is the ratio of the frequency of that class to the total frequency.

Thus in the previous table the relative frequencies of the various classes are $\frac{1}{140}, \frac{6}{140}, \frac{12}{140}, \frac{23}{140}$ etc.

Some times the relative frequency is given as a percentage.

Graphic Representation of a frequency distribution

We have seen that numerical data relating to an event can be presented in the form of a table. As a visual aid to grasp the data in the table, certain diagrams and graphs are used.

The graph representing a frequency distribution is known as a frequency graph. We shall now consider some methods of representing statistical data graphically.

a) The frequency Polygon

Plot the points whose x-co-ordinates are the middle values of the classes and y coordinates are the frequencies in these classes. The figure obtained by joining the successive points is known as a frequency polygon.

b) The Histogram

A second and more important graphical representation of a frequency distribution is by a histogram.

The histogram consists of a set of rectangles erected over the true class intervals, their areas being proportional to the frequencies of the respective classes.

Thus the base of a rectangle is the class-interval in width, the centre of the rectangle is the mid-value and its area represents the class frequency.

Note:

1. In histogram frequencies are represented by areas, where as in a frequency polygon frequencies are represented by lengths.

2. In a histogram the width of a rectangle is the same as that of the class. Since the classes are of equal width, the height of the rectangle will be proportional to the class frequencies.

Frequency table of the marks of table 2 is given below:

Class limits	Mid-value	Frequency
0-9	4.5	1
10-19	14.5	6
20-29	24.5	12
30-39	34.5	23
40-49	44.5	30
50-59	54.5	29
60-69	64.5	19
70-79	74.5	11
80-89	84.5	7
90-99	94.5	2
		140

Measures of Central Tendency

Averages

An average is a value which is typical or representative of a set of data. It represents the whole series and conveys a fairly adequate idea of the whole group. Since such typical values tend to lie centrally within a set of data arranged according to magnitude.

Averages are also called measures of central tendency or measures of location. An average may or may not be one of the variate given in the data.

There are three forms of averages in common use.

The Arithmetic mean

The Median and

The Mode

Averages which are rarely used:

Geometric Mean

Harmonic Mean

1) The Arithmetic Mean or the Mean

i) Individual Observations: The arithmetic mean (A.M) or the mean of a set of n numbers $x_1,x_2,\ldots,x_n$ is denoted by $\bar{x}$ and is defined as

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{r=1}^{n} x_r}{n}$$

ii)

a) Discrete Series: In the case of a frequency distribution, let us that a set of n numbers x1,x2,…,xn having frequencis f1,f2,…,fn respectively.

Then Mean = $\frac{\sum f_r x_r}{\sum f_r}$

ii) Continuous Series: In the case of a frequency distribution, let us that $x_1,x_2,\ldots,x_n$ are the mid values of the class intervals having frequencies $f_1,f_2,\ldots,f_n$ respectively.

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + \ldots + f_nx_n}{f_1 + f_2 + \ldots + f_n} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}$$

Shortcut method for calculating the A.M

$$\bar{x} = A + c\frac{\sum f_r d_r}{N}$$

A- the midpoint of the class interval with the highest frequency

$$d_r = \frac{x_r - A}{c} \qquad N = \sum f_r$$

c- width of the class interval;

Examples:

1) The following table gives the monthly income of 10 employees in an office

1780	1760	1690	1750	1840	1920	1100	1810	1050	1950

Calculate the arithmetic mean of incomes

Solution

Let the income be denoted by the symbol x.

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

$$\frac{1780+1760+1690+1750+1840+1920+1100+1810+1050+1950}{10} = 1665$$

2) From the following data of the marks obtained by 60 students of a class, calculate the arithmetic mean

Marks	No. of students
20	8
30	12
40	20
50	10
60	6
70	4

Solution:

Mean = $\dfrac{\sum f_r x_r}{\sum f_r}$

Marks x_r	No. of students f_r	$f_r x_r$
20	8	160
30	12	360
40	20	800
50	10	500
60	6	360
70	4	280
Total	60	2460

$$\text{Mean} = \frac{\sum f_r x_r}{\sum f_r} = \frac{2460}{60} = 41$$

3) From the following data compute arithmetic mean

Marks	0-10	10-20	20-30	30-40	40-50	50-60
No. of students f_r	5	10	25	30	20	10

Solution:

Class	Mid value x_r	No. of students f_r	$f_r x_r$
0-10	5	5	25
10-20	15	10	150
20-30	25	25	625
30-40	35	30	1050
40-50	45	20	900
50-60	55	10	550
Total		100	3300

$$Mean = \frac{\sum f_r x_r}{\sum f_r} = \frac{3300}{100} = 33$$

Remark: If there is a gap in class interval, find the average of difference, lower-subtraction; upper –add

4) Calculate the A.M by the direct method

Weight in lds	Frequecy	Weight in lds	Frequecy
80-90	1	160-170	9
90-100	11	170-180	5
100-110	25	180-190	4
110-120	37	190-200	3
120-130	62	200-210	2
130-140	31	210-220	1
140-150	22		
150-160	15		

Solution

Class interval	Midvalue of class interval	Frequency	$f_r x_r$
80-90	85	1	85
90-100	95	11	1045
100-110	105	25	2625
110-120	115	37	4255
120-130	125	62	7750
130-140	135	31	4185
140-150	145	22	3190
150-160	155	15	2325
160-170	165	9	1485
170-180	175	5	875
180-190	185	4	740
190-200	195	3	585
200-210	205	2	410
210-220	215	1	215
			29770

$$\bar{x} = \frac{\sum f_r d_r}{\sum f_r} = \frac{29770}{228} = 130.5716$$

5) Calculate the A.M. of frequency distribution of weights of 228 adults given in previous problem by taking an arbitrary origin

Solution

125 is the mid point of the class having the highest frequency. So we take the arbitrary origin A= 125.

Width of the class interval c=10

Class interval	Mid value of class interval x_r	Frequency f_r	$d_r' = \frac{x_r - A}{c}$ $A = 125$	$f_r d_r'$
80-90	85	1	-4	-4
90-100	95	11	-3	-33
100-110	105	25	-2	-50
110-120	115	37	-1	-37
120-130	125	62	0	0
130-140	135	31	1	31
140-150	145	22	2	44
150-160	155	15	3	45
160-170	165	9	4	36
170-180	175	5	5	25
180-190	185	4	6	24
190-200	195	3	7	21
200-210	205	2	8	16
210-220	215	1	9	9
		228		127

$$\bar{x} = A + c\frac{\sum f_r d_r}{N}$$

$$= 125 + 10\left(\frac{127}{228}\right) = 130.570$$

Median

The Median by definition refers to the middle value in a distribution.

Calculation of Mean

i)Individual Observations:

Median is the size of $\frac{N+1}{2}$ item

Discrete Series:

i) Arrange the data in ascending or descending order of magnitude.

ii) Find out the cumulative frequency

iii) Median is the size $\frac{N+1}{2}$ item

iv) Look at the cumulative frequency column and find the total which is either equl to $\frac{N+1}{2}$ or next higher to that and determine the value of the variable corresponding to it. That gives the value of Median.

Continuous Series:

$$\text{Median} = \text{l} + \frac{\text{N/2 - c.f}}{f}.c$$

l- lower limit of the Median class

c.f- cumulative frequency of the class preceeding the median class

f- frequency of the Median class

1) Obtain the value of Median from the following data

391	384	591	407	672	522	777	753	2488	1490

Solution:

Arrange in ascending order

384, 391,407,522,591, 672,753,777,1490,2488

Median = size of $\frac{N+1}{2}$ item = size of (10+1)/2 item

= size of 5.5th item = average of size of 5th item and size of 6th item = (591+672)/2 = 631.5

2) From the following data of the wages of 7 workers , find Median

Wages (Rs.)	1100	1150	1080	1020	1120	1200	1160	1400

Solution:

Median: Arrange the data in ascending order:

1 2 3 4 5 6 7 8

1020, 1080, 1100, 1120, 1150, 1160, 1200, 1400

Median = size of (N+1)/2 item = size of (8+1)/2 =4.5th item = 1135

3) From the following data find the value of Median

Income(Rs.)	1000	1500	800	2000	2500	1800
No. of persons	24	26	16	20	6	30

Solution:

Income	No. of persons f	c.f	
800	16	16	
1000	24	40	
1500	26	66 *	← (N+1)/2=(122+1)/2= 123/2 =61.5
1800	30	96	
2000	20	116	
2500	6	122	

Here N+1/2 = 61.5

In cumulative frequency take which is equal to (N+1)/2 or next heigher to that,

So the c.f which is next higher to N+1/2 is 66.

The corresponding value is 1500

Since Median is the value of variate whose cumulative frequency which is equal to N+1/2 or next heigher

$\therefore$ Median = 1500.

3) calculate the Median for the following distribution

Marks	No. of students
45-50	10
40-45	15
35-40	26
30-35	30
25-30	42
20-25	31
15-20	24
10-15	15
5-10	7

Solution

Marks	f	c.f	
5-10	7	7	
10-15	15	22	
15-20	24	46	
20-25	31	77✓ c.f before the median class	
25-30 median class	42■ f frq. Of m.c	119	← N/2= 200/2 = 100
30-35	30	149	
35-40	26	175	
40-45	15	190	
45-50	10	200	
Total	200		

$$\text{Median} = l + \frac{\text{N/2 - c.f}}{f}.c$$

Median class : 25-30

l-lower limit of the median class =25

c- width of the class interval=5

c.f – cumulative frquency before the median class = 77

f- 42

Median = $25 + \frac{100 - 77}{42}.5$ = 27.7381

From the following data compute Median

Marks	No. of students
410-419	14
420-429	20
430-439	42
440-449	54
450-459	45
460-469	18
470-479	7

There is a gap in the class interval divide the difference by 2 and adjust the limits of class interval

Lower limit – subtract ; upper limit- add

The difference is 1 and divide the difference by 2

½=0.5

Lower limit subtract 0.5; upper limit add o.5

So, the frequency distribution table becomes

Marks	No. of students
409.5-419.5	14
419.5-429.5	20
429.5-439.5	42
439.5-449.5	54
449.5-459.5	45
459.5-469.5	18
469.5-479.5	7

Marks	No. of students	c.f
409.5-419.5	14	14
419.5-429.5	20	34
429.5-439.5	42	76 c.f
439.5-449.5	54 f	130 ←
449.5-459.5	45	175
459.5-469.5	18	193
469.5-479.5	7	200

N/2 =200/2=100

Median class: 439.5-449.5

f=54; c.f= 76; c= 10

$$\text{Median} = l + \frac{\text{N/2 - c.f}}{f}.c$$

$$\text{Median} = 439.5 + \frac{100 - 76}{54}.10 = 443.94$$

Weighted arithmetic Mean:

If $x_1, x_2, \ldots, x_n$ are n observations and $w_1, w_2, \ldots, w_n$ are their respective weights the weighted arithmetic mean is defined to be

$$\text{Weighted A.M} = \frac{w_1x_1 + w_2x_2 + \ldots + w_nx_n}{w_1 + w_2 + \ldots + w_n}.$$

Geometric Mean:

Individual observations: The geometric mean of n sizes $x_1, x_2, \ldots, x_n$ is the nth root of their products

i.e., $(x_1x_2 \ldots x_n)^{1/n}$ (or)

$$\text{G.M.} = \text{Antilog}\left(\frac{\sum \text{logx}}{N}\right)$$

Discrete Series:

$$\text{G.M.} = \frac{1}{N}(f_1 \log x_1 + f_2 \log x_2 + \ldots + f_n \log x_n)$$

Discrete series:

$$\text{G.M.} = \text{Antilog}\left(\frac{\sum \text{f(logx)}}{\sum f}\right)$$

Continuous Series:

$$\text{G.M.} = \text{Antilog}\left(\frac{\sum \text{f(logx)}}{\sum f}\right)$$; x-middle value

Harmonic Mean:

The Harmonic mean of a set of quantities is defined to be the reciprocal of the arithmetic mean of the reciprocal of the quantities. Hence if $x_1, x_2, \ldots, x_n$ are n observations,

$$\text{H.M} = \frac{1}{\frac{1}{\text{n}}\left(\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_n}\right)} = \frac{n}{\sum \frac{1}{x_n}}$$

In a frequency distribution

$$\text{H.M} = \frac{N}{\left(\frac{f_1}{x_1} + \frac{f_2}{x_2} + \ldots + \frac{f_n}{x_n}\right)} = \frac{N}{\sum \frac{f_n}{x_n}}$$

Examples:

1) Calculate Geometric mean from the following data

125	1462	38	7	0.22	0.08	12.75	0.5

Solution:

$$G.M = (x_1 x_2 \ldots x_n)^{1/n} = (54210.3)^{(1/8)} = 6.9517$$

Aliter:

$$\text{G.M.} = \text{Antilog}\left(\frac{\sum \text{logx}}{N}\right)$$

x	logx
125	2.0969
1462	3.1650
38	1.5798
7	0.8451
0.22	-0.657
0.08	-1.096
12.75	1.1055
0.5	-0.3010
Total	6.7367

$$\text{G.M.} = \text{Antilog}\left(\frac{\sum \text{logx}}{N}\right)$$

=6.925

2) Find the geometric mean of the following data:

x	1	2	3	4	5
f	2	4	3	2	1

Find the Geometric mean

Solution:

$$\text{G.M.} = \text{Antilog}\left(\frac{\sum \text{f(logx)}}{\sum f}\right)$$

X	f	logx	f.logx
1	2	0	0
2	4	0.3010	1.204
3	3	0.4771	1.431
4	2	0.6020	1.204
5	1	0.6990	0.6989
Total	12		4.5379

$$\text{G.M.} = \text{Antilog}\left(\frac{\sum \text{f(logx)}}{\sum f}\right) = \text{Antiltog}\,(\frac{4.5379}{12}) = A.L(0.3781) = 2.388$$

3) Find the Geometric mean for the data given below:

Marks	Frequency
4-8	6
8-12	10
12-16	18
16-20	30
20-24	115
24-28	12
28-32	10
32-36	6
36-40	2

Solution

Marks	Mid value x_r	$\log x_r$	f	$f.\log x_r$
4-8	6	0.778	6	4.668
8-12	10	1	10	10
12-16	14	1.146	18	20.628
16-20	18	1.255	30	37.65
20-24	22	1.3424	115	154.376
24-28	26	1.4149	12	16.9788
28-32	30	1.4771	10	14.771
32-36	34	1.5314	6	9.1884
36-40	38	1.579	2	3.158
			209	263.4182

$$\text{G.M.} = \text{Antilog}\left(\frac{\sum \text{f}(\log x_r)}{\sum f}\right)$$

$$G.M. = \text{Antilog}\left(\frac{263.4182}{209}\right) = 18.212$$

Harmonic Mean

1) Calculate the harmonic mean from the following data

3834	382	63	8	0.4	0.03	0.009	0.005

Solution

$$H.M. = \frac{n}{\sum(1/x)}$$

x	1/x
3834	2.608
382	0.003
63	0.016
8	0.125
0.4	2.5
0.03	33.333
0.009	111.111
0.005	200
Total	349.696

$$H.M. = \frac{n}{\sum(1/x)}$$

$$H.M. = \frac{8}{349.696} = 0.023$$

2) From the following data compute the value of harmonic mean

Marks	10	20	25	40	50
No. of students	20	30	50	15	5

Solution

$$H.M. = \frac{N}{\sum(f/x)} \text{ ; N= } \sum f$$

x	f	1/x	f(1/x)
10	20	0.1	2
20	30	0.05	1.5
25	50	0.04	2
40	15	0.025	0.375
50	5	0.02	0.1
Total	120		5.975

$$H.M. = \frac{N}{\sum(f/x)} = \frac{120}{5.975} = 20.08$$

3) From the following data compute the value of harmonic mean

Class	0-10	10-20	20-30	30-40	40-50
Frequency	15	10	7	5	3

$$H.M. = \frac{N}{\sum f(1/x)}$$

X	x_r	f	$1/x_r$	f(1/x)
0-10	5	15	0.2	3
10-20	15	10	0.067	0.67
20-30	25	7	0.04	0.28
30-40	35	5	0.029	0.145
40-50	45	3	0.022	0.066
Total		40		4.091

H.M=40/4.091= 9.778

Measures of Dispersion

Defn: Dispersion is defined as the look of uniformity in the sizes of the items It is the extent to which the value of size differs and hence it is the degree of variability.

The various measures of dispersion are

i) The Range

ii) The Quartile Deviation

iii) The mean deviation

iv) The standard deviation

Range: It is the difference between the greatest and least values observed.

The Quartile Deviation

Quartiles are those values which divide the frequency into four equal parts, where the values are arranged in ascending or descending order of magnitude.

The first Quartile (or the lower quartile) Q_1 is that value of the variate which is such that one quarter of the observations lies below Q_1.

The third quartile (or the upper quartile) Q_3 is that value of the variate which is such that three quarters of the observations lie below Q_3.

The middle or second quartile Q_2 is obviously median. Thus

Individual Observations:

Q_1 is the size of $\frac{N+1}{4}$ th item

Q_3 is the size of $\frac{3(N+1)}{4}$ th item

Discrete and continuous series:

Q_1 is the size of $\frac{N}{4}$ th item

Q_3 is the size of $\frac{3N}{4}$ th item

Continuous Series:

First Quartile:

$$Q_1 = l_1 + \frac{\frac{1}{4}N - c.f_1}{f_1}.c$$

l_1-lower limit of the first quartile class

c-width of the class interval

$c.f._1$= cumulative frequency upto to the lower limit

f_1- frequency of that quartile class

similarly

Third Quartile $Q_3 = l_3 + \frac{\frac{3}{4}N - c.f_3}{f_3}.c$

The Quartile Deviation = $\frac{Q_3 - Q_1}{2}$

The middle or the second quartile Q_2 is the median.

Remark:

Q_1 is that value of x for which cumf = 1/4N

Q_2 is that value of x for which cum f = ½ N

Q_3 is that value of x for which cum f = ¾ N

Example:

1) Calculate the quartile deviation from the following marks of 13 students in a class

25, 35, 9, 28, 52, 41, 38, 96, 85, 72, 10, 40, 60

Soln: Let us arrange the given set of numbers in ascending order

9,10,25,28,35,38,41,41,52,66,72,85,96

Hence N =13

Q_1 is the size corresponding to the rank $\frac{1}{4}(N+1)$ i.e., 3 ½.

Q_1= 25+ 1/2(28-25) = 53/2 (OR)

Q1 = average of 3 rd item and 4th item = $\frac{25+28}{2}=\frac{53}{2}=26.5$

Q_3 is the size corresponding to the rank $\frac{3}{4}(N+1)$. i.e., 10 ½

Q_3= 60+1/2(72-60)=60+6=66

$$Q.D=\frac{Q_3-Q_1}{2}=\frac{66-(53/2)}{2}=\frac{79}{4}$$

2) The following data relate to the frequency distribution of weights of 1000 Males. Calculate the quartile deviation

Wt. in lb	frequency
80.4-94.4	13
94.4-108.4	107
108.4-122.4	340
122.4-136.4	334
136.4-150.4	136
150.4-164.4	48
164.4-178.4	14
178.4-192.4	7
192.4-206.4	-
206.4-220.4	1

Solution:

To find Q_1 and Q_3

We prepare the following cumulative frequency table

Wt. in lb	frequency	Cumulative frequency	
80.4-94.4	13	13	
94.4-108.4	107	120 C.F1	
108.4-122.4 Q1	340 f_1	460 C.F3	¼ N=1/4(930)=232.5
122.4-136.4 Q3	334 f_3	794	¾(N)=3000/4=750
136.4-150.4	136	930	
150.4-164.4	48	978	

164.4-178.4	14	992
178.4-192.4	7	999
192.4-206.4	-	999
206.4-220.4	1	1000

The first Quartile class is 108.4-122.4

$$Q_1 = l_1 + \frac{\frac{1}{4}N - c.f_1}{f_1}.c$$

$$Q_1 = 108.4 + \frac{\frac{1}{4}(1000) - 120}{340}.14 = 113.753$$

The third Quartile class is 122.4-136.4

$$Q_3 = l_3 + \frac{\frac{3}{4}N - c.f_3}{f_3}.c$$

$$Q_3 = 122.4 + \frac{\frac{3}{4}(1000) - 460}{334}.14 = 134.55$$

Hence Quartile Deviation = $Q.D = \frac{Q_3 - Q_1}{2} = \frac{134.55 - 113.753}{2} = 10.3985$

The mode

The value of the variate which occurs most frequently is called mode.

Individual Observations:

Find the mode of the following

4, 7,3,4,8.4

Solution: since 4 appear maximum number of times,

mode = 4

Discrete series:

Find the value of mode from the following data:

Marks:x	No. of students
20	8
30	12

40	20✓
50	10
60	6
70	4

Solution: Since 40 is having highest frequency

Mode = 40

Continuous Series:

Mode= $l + \frac{cf_2}{f_1 + f_2}$

l- lower limit of the modal class; modal class – class with heighest frequency

f1- frequency before the modal class; f2- frequency after the modal class

1) Find the mode in the case of heights of trees in a grade whose frequency distribution is given in the following table

Heights	Frequency
Under 7 feet	26
Under 14 feet	57
Under 21 feet	92-
Under 28 feet	134
Under 35 feet	216
Under 42 feet	287
Under 49 feet	341
Under 56 feet	360

Solution

The given problem becomes

Heights	Frequency
0-7	26
7-14	57-26=31
14-21	92-57=35
21-28	134-92=42 f1
28-35	216-134=82
35-42	287-216=71 f2
42-49	341-287=54
49-56	360-341=19

Since 82 is the largest frequency, the modal class is 28-35

Mode = $l + \frac{cf_2}{f_1 + f_2}$

l- lower limit of the modal class – 28

c-width of the class interval-7

f1- frequency before the modal class-42

f2- frequency after the modal class-71

$= 28 + \frac{7(71)}{42 + 71} = 32.398$

Emperical Relation between mean, median and mode

Mean – Mode = 3(Mean-Median)

1) The following table gives the height of 1000 adult males (measured to the nearest quarter inch):

Height (in inches)	Frequency
58-59.75	2
60-61.75	28
62-63.75	125
64-65.75	270
66-67.75	303
68-69.75	197
70-71.75	65
72-73.75	10

Calculate the mean, median and mode. Verify whether the Emperical relation between them is satisfied.

Solution:

Since there is a gap between the classes we adjust the classes

Difference = 0.25

Divide the difference by 2 i.e., .25/2=.125

Lower limit- subtract 0.125; upper limit add+0.125

So the gn. Problem becomes

Height (in inches)	Frequency
57.875 - 59.875	2
59.875- 61.875	28
61.875-63.75+.125=63.875	125
63.875-65.875	270 f1
65.875-67.875	303✓
67.875-69.875	197 f2
69.875-71.875	65
71.875-73.875	10

I=65.875; c=2; f1= freq. before the modal class=270; f2 = 197

Mode = $l + \frac{cf_2}{f_1 + f_2} = 66.724$

Calculation of mean

Height (in inches)	x_r-mid value	Frequencyf_r	$f_r x_r$
57.875 - 59.875	58.875	2	117.75
59.875- 61.875	60.875	28	1704.5
61.875-63.75+.125=63.875	62.875	125	7859.375
63.875-65.875	64.875	270 f1	17516.25
65.875-67.875	66.875	303✓	20263.125
67.875-69.875	68.875	197 f2	13568.375
69.875-71.875	70.875	65	4606.875
71.875-73.875	72.875	10	728.75
	Total	1000	66365

Mean = $\frac{\sum f_r x_r}{\sum fr} = \frac{66365}{1000} = 66.365$

$$\text{Median} = \text{l} + \frac{\text{N/2 - c.f}}{f}.c$$

N/2= 1000/2 =500

class	Frequencyf_r	Cumulative frequency
57.875 - 59.875	2	2
59.875- 61.875	28	30
61.875-63.75+.125=63.875	125	155
63.875-65.875	270 f1	425 c.f
65.875-67.875	303✓ f	728✓
67.875-69.875	197 f2	925
69.875-71.875	65	990
71.875-73.875	10	1000
Total	1000	

$$\text{Median} = \text{l} + \frac{\text{N/2 - c.f}}{f}.c = 65.875 + \frac{500 - 425}{303}.2 = 66.37$$;

Mean = 66.365; mode = 66.724

The emperical relation between mean, median and mode is

Mean-mode= 3(mean-median)

2mean-3 median= -mode

Mode= 3 median- 2 mean= 3(66.37)-2(66.365)=199.11-132.73=66.38, which is approximately 66.365

Hence emperical relation between mean median mode is satisssfied.

The Mean deviation

The arithmetic mean of the absolute values of the deviations is called the mean deviation or the average deviation.

Individual observation:

$$\frac{\sum \left| x - \bar{x} \right|}{n}$$

In symbols, for a frequency distribution,

Discrete series:

$$\text{Mean deviation} = \frac{\sum f \left| x - \bar{x} \right|}{\sum f}$$

Some times average deviation is also taken from the median.

$$\text{Mean deviation about Median} = \frac{\sum f \left| \text{x - Median} \right|}{\sum f}$$

The standard deviation

The square root of the arithmetic mean of the squares of the deviation is called the standard deviation or the roor mean square deviation.

Individual observations:

$$\sigma = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$$

Discrete series:

$$\sigma = \sqrt{\frac{\sum f(x - \bar{x})^2}{\sum f}}$$

Continuous series:

i) $$\text{Standarddeviation } \sigma = c \cdot \sqrt{\frac{\sum f(x - A)^2}{\sum f}}$$

Here the final value must be multiplied by the width of the class interval to get the value of σ in absolute units.

ii) If the origin A is taken as the arithmetic mean $\bar{x}$, standard deviation is denoted by σ

$$\sigma = c \cdot \sqrt{\frac{\sum f(x - \bar{x})^2}{\sum f}}$$

iii) $$\sigma = c \cdot \sqrt{\frac{\sum \text{fd}^2}{\sum \text{f}} - \left(\frac{\sum \text{fd}}{\sum \text{f}}\right)^2} \quad \text{where} \quad \text{d} = \frac{\text{x - A}}{\text{c}}$$

Relation between the standard deviation and the root square mean square deviation

The root mean square deviation is least when the deviations are measured from the arithmetic mean or the standard deviations is the least possible root mean square deviation.

The square of the standard deviation is also known as the variance of the distribution.

Emperical relation between measures of Dispersion

We have the following approximate relations between the different measure of dispersion

$$\text{Mean Deviation} = \frac{4}{5} \text{Standard Deviation}$$

$$\text{Quartile Deviation} = \frac{2}{3} \text{Standard Deviation}$$

Examples:

1) Find the mean deviation and standard deviation of the heights (in inches) of 16 students given below:

67, 65, 59, 61, 67, 69, 72, 67, 62, 64, 63, 66, 68, 69, 67, 60

Solution

Let us arrange in ascending order

59, 60,61, 62, 63, 64, 64, 65, 66, 67, 67, 67, 67, 68, 69,69, 72

Mean $= \frac{\sum x}{n} = \frac{1046}{16} = 65.375$

X	$x-\bar{x}=x-65.375$	$\left\|x - \bar{x}\right\|$	$\left\|x - \bar{x}\right\|^2$
59	-6.375	6.375	40.641
60	-5.375	5.375	28.891
61	-4.375	4.375	19.141
62	-3.375	3.375	11.391
63	-2.375	2.375	5.641
64	-1.375	1.375	1.891
65	-.375	.375	0.141
66	.625	.625	0.391
67	1.625	1.625	2.641
67	1.625	1.625	2.641
67	1.625	1.625	2.641
67	1.625	1.625	2.641
68	2.625	2.625	6.891
69	3.625	3.625	13.141
69	3.625	3.625	13.141
72	6.625	6.625	43.891
Total		47.25	195.796

$$\text{Mean deviation} = \frac{\sum \left|x - \bar{x}\right|}{n}$$

$$\text{Mean deviation} = \frac{\sum \left|x - \bar{x}\right|}{n} = \frac{47.25}{16} = 2.953$$

$$\text{Standard deviation} = \sqrt{\frac{\sum \left|x - \bar{x}\right|^2}{n}} = \sqrt{\frac{195.756}{16}} = \sqrt{12.23475} = 3.498$$

2) Find the mean deviation and standard deviation for the following distribution

Marks	10	20	25	40	50
No. of	20	30	50	15	5

students					

Solution:

$$\text{Mean deviation} = \frac{\sum f\left|x - \bar{x}\right|}{n}$$

To find $\bar{x}$

$$\bar{x} = \frac{\sum x}{n} = \frac{145}{5} = 29$$

X	$x-\bar{x}=x-29$	$\left\|x-\bar{x}\right\|$	$\left\|x-\bar{x}\right\|^2$	f	$f\left\|x-\bar{x}\right\|$	$f\left\|x-\bar{x}\right\|^2$
10	-19	19	361	20	380	7220
20	-9	9	81	30	270	2430
25	-4	4	16	50	200	800
40	11	11	121	15	165	1815
50	21	21	441	5	105	2205
total				120	1120	14,470

$$\text{Mean deviation} = \frac{\sum f\left|x - \bar{x}\right|}{n} = \frac{1120}{120} = 9.33$$

$$\text{standard deviation} = \sqrt{\frac{\sum f\left|x - \bar{x}\right|^2}{\sum f}} = \sqrt{\frac{14470}{120}} = 10.98$$

3) Compute mean deviation about mean, mean deviation about median and standard deviation for the following distribution

class	f
3.0-4.9	5
5.0-6.9	8
7.0-8.9	30
9.0-10.9	82
11.0-12.9	45
13.0-14.9	24
15.0-16.9	6

Here there is a gap between classes. So we have to adjust the classes

Difference -0.1

Difference/2 = 0.05

Lower limit subtract 0.05 and upper limit add 0.05

So the given problem becomes

class	f
2.95-4.95	5
4.95-6.95	8
6.95-8.95	30
8.95-10.95	82
10.95-12.95	45
12.95-14.95	24
14.95-16.95	6

To find mean deviation

class	Mid value -x	F	fx	$x-\bar{x}$ = x-10.45	$\lvert x-\bar{x}\rvert$	$\lvert x-\bar{x}\rvert^2$	$f\lvert x-\bar{x}\rvert$	$f\lvert x-\bar{x}\rvert^2$
2.95-4.95	3.95	5	19.75	-6.5	6.5	42.25	32.5	211.25
4.95-6.95	5.95	8	47.6	-4.5	4.5	20.25	36	162
6.95-8.95	7.95	30	238.5	-2.5	2.5	6.25	75	187.5
8.95-10.95	9.95	82	815.9	-0.5	0.5	.25	41	20.5
10.95-12.95	11.95	45	537.75	1.5	1.5	2.25	67.5	101.25
12.95-14.95	13.95	24	334.8	3.5	3.5	12.25	84	294
14.95-16.95	15.95	6	95.7	5.5	5.5	30.25	33	181.5
Total		200	2090				369	1158

Mean = $\dfrac{\sum fx}{\sum f} = 10.45$

$$\text{Mean deviation about mean} = \frac{\sum f\left|\text{x} - \bar{\text{x}}\right|}{\sum f} = \frac{369}{200} = 1.845$$

$$\text{Mean deviation about Median} = \frac{\sum f\left|\text{x} - \text{Median}\right|}{\sum f}$$

Median = $l + \dfrac{\frac{1}{2}N - c.f}{f}$

Calculate Median

$$\text{Mean deviation about Median} = \frac{\sum f|\text{x - Median}|}{\sum f} =$$

$$\text{Standarddeviation } \sigma = c \cdot \sqrt{\frac{\sum f(x-\bar{x})^2}{\sum f}} = 2 \cdot \sqrt{5.79} = 4.812$$

Co-efficient of variation:

It is equal to the ratio of the standard deviation of a distribution to it's A.M. and is ofted expressed as a percentage.

$\therefore$

$$\text{coefficient of variation} = \frac{\text{S.D}}{\text{A.M}} \times 100$$

Remark:

For comparing the variability of two series, we calculate the co-efficient of variations for each series.

The series having greater co-efficient of variation is said to more variable or less consistent than the other and the series having lesser co efficient of variation is said to be more consistent or less variable than the other.

Examples: The following table gives 10 measurements of the same quantity under the same conditions by each observer A and B

A	8.116	8.125	8.125	8.129	8.130	8.137	8.137	8.141	8.136	8.146
B	8.112	8.118	8.124	8.130	8.136	8.137	8.138	8.139	8.137	8.141

Calculate the mean value and standard deviation value of each observer's measurements which observe do you think is probably the more reliable and why?

Solution

For measurements of observer A,

$$mean\overline{x_1} = \frac{81.322}{10} = 8.132$$ nearly.

Calculation of standard deviation:

X	$(x-\bar{x})^2$
8.116	0.000256
8.125	0.000049
8.125	0.000049
8.129	0.000009
8.130	0.000004
8.137	0.000025
8.137	0.000025
8.141	.000081
8.136	0.000016
8.146	0.000196

$$\text{S.D} = \sqrt{\frac{\sum(x-\bar{x})^2}{n}} = \frac{0.00071}{10} = \sqrt{0.000071} = 0.008426$$

Hence the co-efficient of variation for A= $\frac{\sigma_1}{\overline{x_1}} \times 100$

$$= \frac{0.008426_1}{8.132} \times 100 = 0.1036$$

For the measurement of observer B,

$$mean\overline{x_2} = \frac{81.312}{10} = 8.1312$$

=8.131 nearly

X	$(x-\bar{x})^2$
8.112	0.00036
8.118	0.000169
8.124	0.000049
8.130	0.000001
8.136	0.000025
8.137	0.000036
8.138	0.000049
8.139	.000064
8.137	0.000036
8.141	0.0001
Total:81.312	0.000889

$$\sigma_2{}^2 = \frac{0.00089}{10} = 0.000089$$

$$\sigma_2 = 0.009434$$

Hence the co-efficient of variation for B = $\frac{\sigma_2}{\bar{x}_2} \times 100 = \frac{0.0009434}{8.131} \times 100 = 0.1160$

The co-efficient of variation in the case of A is smaller than that of B.

A is more reliable than B.

2) The scores of two bats man A and B in a series of matches as follows:

A	37	43	28	62	59	20	83	48	52	47
B	35	52	77	38	26	58	63	31	40	46

Which of the two bats man do you consider the more consistent and more efficient.

X	$(x-\bar{x})^2$	y	$(y-\bar{y})^2$
37	118.81	35	134.56
43	24.01	52	29.16
28	396.01	77	924.16
62	198.81	38	73.96
59	123.21	26	424.36
20	728.41	58	129.96
83	1232.01	63	268.96
48	.01	31	243.36
52	16.81	40	43.56
47	0.81	46	.36

$\bar{x} = \frac{\sum x}{n} = 47.9$; $\bar{y} = \frac{\sum y}{n} = 46.6$

$\sigma_1^2 = \frac{2888.9}{10} = 288.89$

$\sigma_1 = 16.996 = 17$ nearly

$\sigma_2^2 = \frac{2272.4}{10} = 227.24$

$\sigma_2 = 15.07$

$\therefore$

The coefficient of variation for A = $\frac{\sigma_1}{x_1} \times 100 = \frac{17}{47.9} \times 100$ 35.49%

The coefficient of variation for B = $\frac{\sigma_2}{x_2} \times 100 = \frac{15.07}{46.6} \times 100$ 32.34%

Since the A.M. of A is greater than the A.M of B, we conclude that A is more efficient than B.

Since the coefficient of variation B is less than the co efficient of variation of A, we conclude that B is more consistent than A.

Thus even though A is the better player , he is less consistent.

Exercises:

1) The score of two golfers A and B in 12 rounds are given below. Who is the better player and who is the more consistent player

A	74	75	78	72	78	77	79	81	79	76	72	71
B	87	84	80	88	89	85	86	82	82	79	86	80

To find the S.D of the combination of two groups

Let n_1, $\overline{x_1}$ and σ_1 be the frequency, the A.M. and S.D. of a first set of variables and those respectively for the second set be denoted by n_2, $\overline{x_2}$ and σ_2. Let $\bar{x}$ be the A.M. of the combined set of $n_1 + n_2$ variables.

If $n_1 + n_2 = N$ then $N\sigma^2 = n_1\sigma_1^2 + n_2\sigma_2^2 + n_1D_1^2 + n_2D_2^2$

Where $D_1 = \overline{x_1} - \bar{x}$, $D_2 = \overline{x_2} - \bar{x}$

Examples:

1) The numbers examined, the mean weight and standard deviation in each group of the examinations by three medical examiners are given below. Find the mean weight and standard deviation of the entire data when grouped together.

Medical Examiner	No. Examined	Mean weight Kg.	S.D. kg.
A	50	56	3
B	60	60	4
C	90	58	5

If σ is the standard deviation of the entire data then

$$N\sigma^2 = n_1\sigma_1^2 + n_2\sigma_2^2 + n_3\sigma_3^2 + n_1D_1^2 + n_2D_2^2 + n_3D_3^2$$

Where $D_1 = \overline{x_1} - \overline{x}$, $D_2 = \overline{x_2} - \overline{x}$, $D_3 = \overline{x_3} - \overline{x}$

and $N= n_1+n_2+n_3$

$D_1= 56-58.1=-2.1$

$D_2= 60-58.1= 1.9$

$D_3=58-58.1=-.1$

$N= n_1+n_2+n_3=200$

$$N\sigma^2 = 50(3)^2 + 60(4)^2 + 90(5)^2 + 50(-2..1)^2 + 60(1.9)^2 + 90(-0.1)^2$$

$$N\sigma^2 = 4098$$

i.e., $200\sigma^2 = 4098$

$$\sigma^2 = 20.49$$

$$\sigma = 4.527 \text{ k.g}$$

Exercises:

1) Find the mean and S.D. of the following two samples put together

Sample No.	Size	Mean	S.D
1	50	158	5.1
2	60	164	4.6

UNIT IV

Linear Correlation

Correlation coefficient

Correlation:

The existence of the changes in one variable in sympathy with the changes in the other is called correlation.

Thus, whenever two variables are related in such a way that a change in one is followed directly or inversely by a change in the other they are said to be correlated.

The Scatter Diagram

In the case of raw correlated data, we can represent them graphically. Let (x_1, y_1), (x_2,y_2),…, (x_n,y_n) be pairs of corresponding observations .

For example, $x_1, x_2,\ldots,x_n$ may be the ages of husbands and $y_1, y_2,\ldots$, yn may be the ages of wives. Plot the points(x_1,y_1), (x_2, y_2) etc on a graph paer. The figure which ijs simply a collection of dots is called the dot diagram or the scatter diagram. From this scatter diagram we can guess roughly how the variables x and y are correlated.

If all the points in the scatter diagram seem to lie near a line as in figure 1(a), there is correlation between the variables and the correlation is called linear.

If all the points seem to cluster round some curve as in figure 1 (b), the cojrrelation is called the non linesr.

If the amount of change in oe variable tens to bear constant ratio to the amount of change in the other variable then the correlation is said to be linear.

If the amount of change in one variable does not bear a constant ratio to the amount of change in the other variable then the correlation is called nonlinear or curvilinear.

Coefficient of Correlation

Let $\bar{x}$ and $\bar{y}$ be respective arithmetic means of x1,x2,…xn and y1, y2, …, yn. There is said to be positive correlation between x and y if, for any assigned value of $x>\bar{x}$, the correspondingvalues of y tend to be $>\bar{y}$ and if for any assigned x less than $\bar{x}$ the corresponding y values tend to be $<\bar{y}$.

The correlation is said to be negative if for $x>\bar{x}$, y tends to be $<\bar{y}$ and if for $x<\bar{x}$, y tends to be $>\bar{y}$

The quantity $P = \dfrac{\sum\left(x - \bar{x}\right)\left(y - \bar{y}\right)}{N}$ is said to be the covariance between x and y.

$$r = \frac{P}{\sigma_x \sigma_y} = \frac{\sum (x-\bar{x})(y-\bar{y})}{N\sigma_x\sigma_y} = \frac{\sum (x-\bar{x})(y-\bar{y})}{N\sqrt{\frac{\sum (x-\bar{x})^2}{N}}\sqrt{\frac{\sum (y-\bar{y})^2}{N}}} = \frac{\sum (x-\bar{x})(y-\bar{y})}{\frac{N\sqrt{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2}}{\sqrt{N^2}}}$$

r is called Pearson's product moment correlation coefficient or correlation co-efficient.

Computation of r –Direct Method

$$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{N\sigma_x\sigma_y}$$

$$\sigma_x = \sqrt{\frac{\sum (x-\bar{x})^2}{N}} \; ; \quad \sigma_y = \sqrt{\frac{\sum (y-\bar{y})^2}{N}}$$

Given a set of values of x and y, we can calculate $\bar{x}$ and $\bar{y}$ by using the formula

$x - \bar{x} = X \qquad y - \bar{y} = Y$

If and then $\sigma_x = \sqrt{\frac{\sum X^2}{N}}$, $\sigma_y = \sqrt{\frac{\sum Y^2}{N}}$

$$r = \frac{\sum XY}{N\sigma_x\sigma_y} = \frac{\sum XY}{\sqrt{\sum X^2 Y^2}}$$

Remark:

The correlation coefficient always lies between -1 and 1.

when r = -1 , it means that there is perfect negative correlation between the variables.

When r = +1, it means that there is perfect positive correlation between the variables.

When r = 0, it means that there is no relationship between the two variables.

Examples:

1) Calculate the Karl Pearson's coefficient of correlation from the following data

Roll No. of students	1	2	3	4	5
Mark in Accountancy	48	35	17	23	47
Marks in Statistics	45	20	40	25	45

Solution

Let the Marks in Accountancy be denoted by X and Marks in Statistics be denoted by Y.

Roll No.	X	$X - \bar{X}$	$(X-\bar{X})^2$	Y	$Y - \bar{Y}$	$(Y-\bar{Y})^2$	$(X-\bar{X})(Y-\bar{Y})$
		$X - 34$			$Y - 35$		

	48	14	196	45	10	100	140
	35	1	1	20	-15	225	-15
	17	-16	256	40	5	25	-80
	23	-11	121	25	-10	100	110
	47	13	169	45	10	100	130
Total	170		743	175		550	285

$$\overline{X} = \frac{\sum x}{n} = \frac{170}{5} = 34 \; ; \; \bar{y} = \frac{\sum y}{n} = \frac{175}{5} = 35$$

$$r = \frac{\sum \left(X - \overline{X}\right)\left(Y - \overline{Y}\right)}{\sqrt{\sum \left(X - \overline{X}\right)^2 \sum \left(Y - \overline{Y}\right)^2}}$$

$$= \frac{285}{\sqrt{(743)(550)}} = \frac{285}{639.257} = 0.446$$

2) Calculate coefficient of correlation from the following data

X	100	200	300	400	500	600	700
Y	30	50	60	80	100	110	130

Solution

To simplify calculation let every value of X be divided by 100 and every value of Y be divided by 10 and denote these series by X' and Y'

X	X'=X/100	$X - \overline{X}'$	$\left(X - \overline{X}'\right)^2$	Y	Y'=Y/10	$Y - \overline{Y}'$	$\left(Y - \overline{Y}'\right)^2$	
100	1	-3	9	30	3	-5	25	15
200	2	-2	4	50	5	-3	9	6
300	3	-1	1	60	6	-2	4	2
400	4	0	0	80	8	0	0	0
500	5	1	1	100	10	2	4	2
600	6	2	4	110	11	3	9	6
700	7	3	9	130	13	5	25	15
	28 $\overline{X}' = 4$		28		56 $\overline{Y}' = 8$		76	46

$$r = \frac{\sum \left(X' - \overline{X}\right)\left(Y' - \overline{Y}'\right)}{\sum \left(X' - \overline{X}'\right)^2 \sum \left(Y' - \overline{Y}'\right)^2}$$

$$r = \frac{(46)}{\sqrt{(28)(76)}} = \frac{46}{46.130} = 0.997$$

3) Find the correlation coefficient between the heights of father and heights of son given below:

Height of father (in inches)	65	66	67	67	69	71	72	70	65
Height of Son (in inches)	67	68	69	68	70	70	69	70	70

Solution

Let x and Y be the heights of father and heights of son respectively

$$\bar{x} = \frac{\sum x}{n} = \frac{612}{9} = 68 \; ; \; \bar{y} = \frac{\sum y}{n} = \frac{621}{9} = 69$$

X	Y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(y - \bar{y})^2$	$(x - \bar{x})(y - \bar{y})$
65	67	-3	-2	9	4	6
66	68	-2	-1	4	1	2
67	69	-1	0	1	0	0
67	68	-1	-1	1	1	1
69	70	1	1	1	1	1
71	70	3	1	9	1	3
72	69	4	0	16	0	0
70	70	2	1	4	1	2
65	70	-3	1	9	1	-3
Total:612	621			54	10	12

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}} = \frac{12}{\sqrt{(54)(10)}} = \frac{12}{23.238} = 0.5164$$

Computation of r –Shortcut

Method:

$$p = \frac{\sum d_1 d_2}{N} - \frac{\sum d_1}{N} \cdot \frac{\sum d_2}{N}$$

$$\sigma_x{}^2 = \frac{\sum d_1{}^2}{N} - \left(\frac{\sum d_1}{N} \right)^2$$

$$\sigma_y^2 = \frac{\sum d_2^2}{N} - \left(\frac{\sum d_2}{N}\right)^2$$

$$r = \frac{P}{\sigma_x \sigma_y}$$

4) Find the co-efficient of correlation between industrial production and export using the following data and comment on the result

Product (in crore tons)	55	56	58	59	60	60	62
Exports (in crore tons)	35	38	38	39	44	43	44

Solution:

Let x represent the product and y represent the export

Take the origin at 59 for x and 39 for y. We prepare the following table.

X	Y	d_1=x-A	d_2=Y-B	d_1^2	d_2^2	d_1d_2
55	35	-4	-4	16	16	16
56	38	-3	-1	9	1	3
58	38	-1	-1	1	1	1
59	39	0	0	0	0	0
60	44	1	5	1	25	5
60	43	1	4	1	16	4
62	44	3	5	9	25	15
		-3	8	37	84	44

$$\mathrm{p} = \frac{\sum \mathrm{d_1 d_2}}{\mathrm{N}} - \left(\frac{\sum \mathrm{d_1}}{\mathrm{N}}\right)\left(\frac{\sum \mathrm{d_2}}{\mathrm{N}}\right)$$

$$\mathrm{p} = \frac{44}{7} - \left(\frac{-3}{7}\right)\left(\frac{8}{7}\right)$$

$$\mathrm{p} = \frac{308 + 24}{49} = \frac{332}{49}$$

$$\sigma_x^2 = \frac{\sum d_1^2}{N} - \left(\frac{\sum d_1}{N}\right)^2$$

$$\sigma_x{}^2 = \frac{37}{7} - \left(\frac{-3}{7}\right)^2$$

$$\sigma_x{}^2 = \frac{259-9}{49} = \frac{250}{49}$$

$$\sigma_y{}^2 = \frac{\sum d_2{}^2}{N} - \left(\frac{\sum d_2}{N}\right)^2$$

$$\sigma_y{}^2 = \frac{84}{7} - \left(\frac{8}{7}\right)^2$$

$$\sigma_y{}^2 = \frac{588-64}{49} = \frac{524}{49}$$

$$\therefore r = \frac{p}{\sigma_x \sigma_y} = \frac{332/49}{\sqrt{250/49}\sqrt{524/49}}$$

$$= \frac{332}{\sqrt{250}\sqrt{524}} = \frac{332}{361.9392} = 0.9173$$

Exercises:

1)Calculate the pearson's coefficient of correlation form the following data using 44 and 26 respectively as the origin of x and y (instead of x bar use 44 and instead of y bar use 26 in the formula)

X	43	44	46	40	44	42	45	42	38	40	42	57
Y	29	31	19	18	19	27	27	29	41	30	26	10

(r=-0.733) $$r = \frac{\sum (x-A)(y-B)}{\sqrt{\sum (x-A)^2 \sum (y-B)^2}} = \frac{\sum (x-44)(y-26)}{\sqrt{\sum (x-44)^2 \sum (y-26)^2}}$$

2)From the following table calculate the coefficient of correlation by Karl Pearson's method

X	6	2	10	4	8
Y	9	11	?	8	7

Arithmetic mean of x and y are 6 and 8 respectively

Solution:

From the mean value of y we can find the missing value

(9+11+ _+8+7)/5=8 (gn)

Missing value = 40-(35)=5

(r=-.919)

3) The following table gives indices of industrial production of registered unemployed (in hundred thousand). Calculate the value of the coefficient so obtained.

Year	1991	1992	1993	1994	1995	1996	1997	1998
Index of Production	100	102	104	107	105	112	103	99
Number unemployed	15	12	13	11	12	12	19	26

Regression Lines

Let (x_1, y_1) , $(x_2, y_2),\ldots,(x_n,y_n)$ be n observations of the two variables. If we plot these n points on a graph paper, it may happened that these points tend to cluster themselves along sojme well defined lines. These are called regression lines.

Regression Equations

Regression equations also known as estimating equations are algebraic expressions of the regression lines, there are two regression equations.

The regression equation of X on Y is used to describe the variations in the values of X for given changes in Y and the regression equations of Y on X is used to describe the variation in the values of y for given changes in X.

Regression equation of Y on X

$$Y - \overline{Y} = r\frac{\sigma_y}{\sigma_x}\left(X - \overline{X}\right)$$

Regression equation of X on Y

$$X - \overline{X} = r\frac{\sigma_x}{\sigma_y}\left(Y - \overline{Y}\right)$$

Remark:

1. Regression line of y on x is used to find the probable value or expected value of y for a given value of x.
2. The regression line of x on y is used to find the probable or expected value of x for a given value of y.
3. Both the regression lines passing through $(\overline{x}, \overline{y})$
4. The quantities are called the regression coefficients.
5. If the regression coefficients are both positive, then r is positive. If the regression coefficients are both negative then r is negative.

6. $r = \frac{p}{\sigma_x \sigma_y}$; $r^2 = \frac{p^2}{\sigma_x{}^2 \sigma_{y^2}} = \frac{p}{\sigma_x{}^2} \cdot \frac{p}{\sigma_y{}^2}$

Hence r is the G.M. between the two regression coefficients.

7. Since $r = \frac{p}{\sigma_x \sigma_y}$, $\frac{p}{\sigma_{x^2}} = r \cdot \frac{\sigma_y}{\sigma_x}$ and $\frac{p}{\sigma_y{}^2} = r \cdot \frac{\sigma_x}{\sigma_y}$

Hence the regression lines have the equation

$y - \bar{y} = r \frac{\sigma_y}{\sigma_x}(x - \bar{x})$ and $x - \bar{x} = r \frac{\sigma_x}{\sigma_y}(y - \bar{y})$

Note:

1) $b_{yx} = r \frac{\sigma_x}{\sigma_y}$ is the regression co-efficient of X on Y.

2) $b_{xy} = r \frac{\sigma_y}{\sigma_x}$ is the regression coefficient of Y on X.

3) $r = \sqrt{b_{xy} \times b_{yx}}$

1) The following data relate to the scores obtained by 9 salesmen of a company in an intelligence test and their weekly sales in thousand rupees.

Test Scores	50	60	50	60	80	50	80	40	70
Weekly Sales	30	60	40	50	60	30	70	50	60

a) Obtain the set regression equation of sales on intelligence test scores of the salesmen.

b) Obtain the intelligence test score of a salesman is 64. What would be his expected weekly sales.

2) The following regression equations were obtained from a correlation table:

$y = 0.516x + 33.73$

$x = 0.312y + 32.52$

Find the value of

i) the correlation coefficient

ii) the mean of x

iii) the mean of y

Solution

Since the equations to the regression lines are

$y-\bar{y}=\frac{p}{\sigma_x{}^2}(x-\bar{x})$ and $x-\bar{x}=\frac{p}{\sigma_y{}^2}(y-\bar{y})$

These are the regression coefficients and they are both positive.

Hence the correlation is direct.

Correlation coefficient = $\sqrt{\frac{p}{\sigma_x{}^2}\cdot\frac{p}{\sigma_y{}^2}}$ (since regression coefficients are given, r= square root of mult. (G.M.)Of both the regression coefficients)

$=\sqrt{(0.516)(0.512)}=0.514$

$(\bar{x},\bar{y})$ is the point of intersection of the regression lines.

$y=0.516x+33.73$

$x=0.312y+32.52$

Solving (1) and (2)

$-0.516x+y=33.73$ ----------(1)

$(2)\times 0.516 \Rightarrow 0.516x - 0.1601y = 16.780$.....(3)

$(1)+(3) \Rightarrow 0.8399y=16.95$

$y=20.181$

-0.516x+y=33.73 -----(1)

0.516x-.264y=16.78 ___(2)

_____

0.736y=50.51

Y = 68.63

Substitute in (1)

X= 67.64

Hence mean is (67.64, 68.63)

3) From the following data, find the most likely value of y when x = 24

	Y	x
Mean	958.8	18.1

S.D	36.4	2.0

Also r=0.58

Solution

The equation to the regression line of y on x is

$$y-\bar{y}=r\frac{\sigma_y}{\sigma_x}\left(x-\bar{x}\right)$$

$$y-958.8=(0.58)\frac{36.4}{2}\left(x-18.1\right)$$

y-958.8=10.556x-191.064

$$y=10.556x+767.736$$

Putting x =24

Y= 253.344+767.736=1021.08

Exercises:

4) Find the equation of regression lines for the data given below:

X	25	28	35	32	36	36	29	38	34	22
Y	43	46	49	41	36	32	31	30	33	39

5) Find the regression line of y on x

X	1	2	3	4	5	6	7	8	9
Y	9	8	10	12	11	13	14	16	15

Also obtain an estimate of y which should correspond on an average to x =6.2

Rank Correlation:

$$\rho=1-\frac{6\sum d^2}{n(n^2-1)}$$ where d= X-Y

Exercise:

Obtain the rank correlation between the variables X and Y from the following pairs of observed values

X	50	55	65	50	55	60	50	65	70	75
Y	110	110	115	125	140	115	130	120	115	160

UNIT-IV

Fitting a straight line by the method of least squares

Let (xi, i) , I = 1,2,…,n be the n sets of observations

Let $y= a_0 +a_1x$ be the best fir to the data

The normal equations are

$$na_0 + a_1 \sum x = \sum y$$

$$a_0 \sum x + a_1 \sum x^2 = \sum xy$$

The solutions are

$$a_0 = \frac{\sum y \sum x^2 - \sum x \sum xy}{n\sum x^2 - \left(\sum x\right)^2}$$

$$a_1 = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2}$$

Examples:

1) Fit a straight line for the table of values given below:

Indept. Variable x	1	2	4	5	6	8	9
Dependent Variable y	2	5	7	10	12	15	19

Solution

Let the straight line be

$Y= a_1x+a_0$

The normal equations are

$$na_0 + a_1 \sum x = \sum y$$

$$a_0 \sum x + a_1 \sum x^2 = \sum xy$$

X	y	X^2	xy
1	2	1	2
2	5	4	10
4	7	16	28

5	10	25	50
6	12	36	72
8	15	64	120
9	19	81	171
Total:35	**70**	**227**	**453**

$$a_0 = \frac{\sum y \sum x^2 - \sum x \sum xy}{n\sum x^2 - \left(\sum x\right)^2}$$

$$a_0 = \frac{(70)(227) - (35)(453)}{7(227) - (35^2} = \frac{35}{364} = 0.0962$$

=- 43.4615

$$a_1 = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2}$$

$$a_1 = \frac{7(453) - (35)(70)}{7(227) - (35)^2} = \frac{721}{364} = 1.9808$$

Y=1.9808x+43.4615

Remark:

If for the same table of values we assume y to be the independent variable and x to be the dependent variable, we consider an equation of the type

$X= a_1y+a_0$

In this case

$$a_0 = \frac{\sum y \sum y^2 - \sum y \sum xy}{n\sum y^2 - \left(\sum y\right)^2}$$

$$a_1 = \frac{n\sum xy - \sum x \sum y}{n\sum y^2 - \left(\sum y\right)^2}$$

Polynomial Regression

Assume that n pairs of coordinates (x_1,y_1) are given which are to be approximated by a quadratic. Let the quadratic curve be represented by

$Y=a_2x^2+a_1x+a_0$

The normal equations are

$$na_0 + a_1\sum x + a_2\sum x^2 = \sum y$$
$$a_0\sum x + a_1\sum x^2 + a_2\sum x^3 = \sum xy$$
$$a_0\sum x^2 + a_1\sum x^3 + a_2\sum x^4 = \sum x^2 y$$

Examples:

1) Fit the quadratic curve to the following data:

X	**-4**	**-3**	**-2**	**-1**	**0**	**1**	**2**	**3**	**4**	**5**
Y	**21**	**12**	**4**	**1**	**2**	**7**	**15**	**30**	**45**	**67**

Solution

Let $y= a_2x^2+a_1x+a_0$

The normal equations are

$$na_0 + a_1\sum x + a_2\sum x^2 = \sum y$$
$$a_0\sum x + a_1\sum x^2 + a_2\sum x^3 = \sum xy$$
$$a_0\sum x^2 + a_1\sum x^3 + a_2\sum x^4 = \sum x^2 y$$

X	Y	X^2	X^3	X^4	xy	X^2y
-4	21	16	-64	256	-84	336
-3	12	9	-27	81	-36	108
-2	4	4	-8	16	-8	16
-1	1	1	-1	1	-1	1
0	2	0	0	0	0	0
1	7	1	1	1	7	7
2	15	4	8	16	30	60
3	30	9	27	81	90	270
4	45	16	64	256	180	720
5	67	25	125	625	335	1675

The normal equations are

$$10a_0 + 5a_1 + 85a_2 = 204$$
$$5a_0 + 85a_1 + 125a_2 = 513$$
$$85a_0 + 125a_1 + 1333a_2 = 3193$$

Solving these, we get

$a_2=1.98$; $a_1= 3$, $a_0=2.07$

Fitting Exponential and Trigonometric equations:

Let $y = ae^{bx}$ be the curve to be fitted.

Taking log on both sides

$$\log y = \log_{10} a + bx \log_{10} e$$

Y=A+Bx where A= $\log_{10}$ a; B= b $\log_{10}$e

The normal equations are

$$\sum Y = nA + B\sum x$$
$$\sum xY = A\sum x + B\sum x^2$$

Examples:

1) Fit a curve $y = ae^{bx}$ to the above data:

x	0	5	8	12	20
y	3	1.5	1	0.55	0.18

x	Y	Y=log $_{10}$y	x^2	xY
0	3	0.4771	0	0
5	1.5	0.1761	25	0.8805
8	1	0	64	0
12	0.55	-0.2596	144	-3.1152
20	0.18	-0.7447	400	-14.894
		-0.3511	633	-17.1287

A=0.4815; B=-0.0613

Log $_{10}$a = 0.4815

a=10^ (0.4815)= Anti log(0,4815)= 3.0304

B=-0.0613

B=-0.0613/log $_{10}$e

= -0.0613/ 0.4343

= -0.1411

y=3.0304 $e^{-0.1411x}$

Examples:

1) Fit a curve $y = ax^b$ to the following data:

x	1	2	3	4	5	6
y	151	100	61	50	20	8

Solution:

$Y= ax^b$

Taking $\log_{10}$

$\text{Log}_{10}y=\log_{10} a+x\log_{10}b$

$Y=A+Bx$ where $B= \log_{10}b$

The normal equations are

$$\sum y = nA + B\sum x$$
$$\sum xY = A\sum x + B\sum x^2$$

x	Y	Y=logy	xY	X^2
1	151	2.1790	2.1790	1
2	100	2	4	4
3	61	1.7853	5.3559	9
4	50	1.6990	6.796	16
5	20	1.3010	6.505	25
6	8	0.9031	5.4186	36
		9.8674	30.2545	91

The normal equations are

$$\sum y = nA + B\sum x$$
$$\sum xY = A\sum x + B\sum x^2$$

$$9.8674 = 6A + 21B$$
$$30.3545 = 21A + 91B$$

Solving these

A= 2,5010, B= -0.2447

$\text{Log}_{10}a=2.5010$

a=10^ 2.5010=316.9567

b=10^B=0.5692

$$y = 316.9567(0.5692)^x$$

3) Fit a curve $y= ax^b$ to the following data:

X	1	2	3	4	5	6
Y	1200	900	600	200	110	50

Solution:

$Y=ax^b$

$\log_{10} y = \log_{10} a + b \log_{10} x$

$Y = A + BX$

$A = 3.3086$

$b = -1.7494$

$y = 2035x^{-1.7494}$

LARGE SAMPLES

Definition

The group of individuals under study is called population or universe.

The population may be finite or infinite

Definition

A finite subset of statistical individuals in a population is called sample.

Definition

The number of individuals in a sample is called sample size.

Example

In a shop we asses the quality of rice or any other commodity by taking a handful of it from the bag and decide to purchase it or not.

Parameters and statistics

The statistical constants of the population namely mean μ, variance σ^2 which are usually referred to as parameters.

Statistical measures computed from sample observations alone

Example: Mean, variance etc are usually referred to as statistic.

Sampling Distribution

If we draw a sample of size n from a given finite population of size N then the total number of possible samples is N_{Cn}.

$$N_{C_n} = \frac{N!}{n!(N-n)!} = K$$

For each of these k samples we can compute, some statistic say $t = t(x_1.x_2\ldots,x_n)$ in particular the mean, variance etc as given below:

The set of values of the statistic so obtained, one for each sample constitutes the sampling distribution of the statistic.

Standard error:

The standard deviation of sampling distribution of a statistic is known as its standard error and it is denoted by (S.E.)

Test of Significance:

A very important aspect of the sampling theory is the study of tests of significance which en enable us to decide on the basis of the sample results if

i) The deviation between the observed sample statistic and the hypothetical parameter value is significant.

ii) The deviation between two sample statistics is significant.

Null hypothesis:

For applying the test of significance we first set up of a hypothesis a definite statement about the population parameter. Such a hypothesis is usually a hypothesis of no difference and it is denoted by H_0.

Alternative hypothesis:

Any hypothesis which is complementary to the null hypothesis is called an alternative hypothesis, usually denoted by H_1.

For example

If we want to test the null hypothesis that the population has a specified mean μ_0 (say) i.e., H_0: $\mu = \mu_0$

Then the alternative hypothesis would be (i) H_1: $\mu \neq \mu_0$ (i.e., either $\mu > \mu_0$ or $\mu < \mu_0$)

ii) H_1: $\mu > \mu_0$

iii) H_1: $\mu < \mu_0$

The alternative hypothesis (i) is known as a two tailed alternative and the alternative in (ii) is known as right tailed and (iii) is known as left tailed.

The setting of alternative hypothesis is very important to decide whether we have to use a single tailed (right or left) or two tailed test.

Errors in Sampling

The main objective in a sampling theory is to draw valid inferences about the population parameters on the basis of the sample results. In practice we decide to accept or reject the lot after examining a sample from it. We have two type of errors.

Type I error: Reject H0 when it is true.

Type II error: Accept H_0 when it is wrong

Critical region

A region corresponding to a statistic t in the sample space S which lead to the rejection of H_0 is called critical region or Rejection region. Those region which lead to the acceptance of $_{H0}$ give us a region called acceptance region.

Level of Significance

The probability α that a random value of the statistic t belongs to the critical region is known as the level of significance. In otherwords, level of significance is the size of the Type I error. The levels of significance usually employed in testing of hypothesis are 5% and 1%.

One tailed and Two tailed test:

A test of any statistical hypothesis where the alternative hypothesis is one tailed (right tailed or left tailed) is called one tailed test.

In a test of statistical hypothesis where the alternative hypothesis is two tailed is called two tailed test.

Procedure for testing of hypothesis:

i) Set up the null hypothesis

ii) Choose the appropriate level of significance (either 5% or 1%) This is to be decided before sample is drawn.

iii) Compute the test statistic $z = \dfrac{t - E(t)}{S.E(t)}$ under the null hypothesis

iv) we compare the computed value of z in step (iii) with the significant value at given level of significance.

If $|z| < 1.96$ H_0 may be accepted at 5% level of significance.

If $|z| > 1.96$ H_0 may be rejected at 5% level of significance.

$|z| < 2.58$ H_0 may be accepted at 1% level of significance

$|z| > 2.58$ H_0 may be rejected at 1% level of significance.

For single tailed test (Right tail or left tail) we compare the computed value of z with 1.645 (at 5% level) and 2.33 (at 1% level) and accept or reject H_0 accordingly.

Remark:

	Two tailed	One-tailed

5%	1.96	1.645
1%	2.58	2.33

Calculated value <table value - accept the null hypothesis

Calculated value > table value -reject the null hypothesis

Large Samples:

Definition

If the size of the sample n>30 then that sample is called large sample.

Test of Significance of large samples:

There are 4 important test to test the significance of large samples.

1. Test of significance for single proportion

2. Test of significance for difference of Proportions

3. Test of significance for single mean

4. Test of significance for difference of means

1. Test of significance for single Proportion

Suppose a large sample of size n is taken from a normal population. To test the significant difference between the sample proportion p and the population proportion P , we use the test statistic

$$z = \frac{p - P}{\sqrt{\frac{PQ}{n}}}$$ where n- sample size

Note: Limits for population P are given by $p \pm 3\sqrt{\frac{PQ}{n}}$ where q=1-p

Examples:

1) A manufacturer claimed that atleast 95% of the equipment which he supplied to a factory conformed to specifications. An examination of a sample of 200 pieces of equipment revealed that 18 were faulty. Test his claim at 5% level of significance?

Solution

Given sample size n = 200

Number of pieces confirming to specification = 200-18=182

$\therefore$

p= proportion of pieces confirming to specifications = $\frac{182}{200} = .91$

P= Population proportion = $\frac{95}{100}$

Null hypothesis H_0 : The proportion of pieces confirming to specification

i.e., P=95%

Alternative hypothesis: H_1: P<.95 (one-tailed alternative)

Test statistic $z = \frac{p - P}{\sqrt{\frac{PQ}{n}}}$

$= \frac{0.91 - 0.95}{\sqrt{\frac{.95 \times .05}{200}}}$

= -2.59

Since alternative hypothesis is one tailed, the tabulated value of z at 5% level of significance is 1.645

Since calculated value of $|z| = 2.6$ is >1.645, we reject the null hypothesis H_0 at 5% level of significance.

Hence the manufacturer claim is rejected.

2) In a sample of 1000 people in Karnataka 540 are rice eaters and the rest are wheat eaters. Can we assume that both rice and wheat are equally popular in this state at 1% level of significance?

Solution

Given n =1000

P= sample proportion of rice eaters

$= \frac{540}{1000} = .54$

P= Population proportion of rice eaters

$= \frac{1}{2} = .5$

Q=1-P=0.5

Null hypothesis: H_0: Both rice and wheat are equally popular in the state i.e., p=P orP= 1/2

Alternative hypothesis: H_1: $P \neq 0.5$ (Two tailed alternative) or $p \neq P$

Test statistic

$$z = \frac{p - P}{\sqrt{\frac{PQ}{n}}}$$

$$z = \frac{0.54 - 0.5}{\sqrt{\frac{(0.5)(0.5)}{1000}}}$$

=2.532

The tabulated value of z at 1% level of significance is 2.58 for two tailed test.

Since calculated value of z< tabulated value of z, we accept H_0.

i.e., Both rice and wheat eaters are equally popular in that state.

3) In a sample of 400 parts manufactured by a factory, the number of defective parts was found to be 30. The company, however claimed that only 5% of their product is defective. Is the claim tenable?

Solution

Given n = 400

No. of defectives in the sample = 30

p= proportion of defectives in the sample = $\frac{30}{400} = 0.075$

P= the population proportion = $\frac{5}{100} = 0.05$

$\therefore Q = 1 - P = 1 - 0.05 = 0.95$

Null hypothesis H_0 : The company's claim P =0.05 is acceptable.

Alternative hypothesis H_1: P>0.05 (one tailed alternative)

Test statistic

$$z = \frac{p - P}{\sqrt{\frac{PQ}{n}}}$$

$$z = \frac{0.075 - 0.050}{\sqrt{\frac{0.05 \times 0.95}{400}}}$$

=2.27

Since the alternative hypothesis is one tailed alternative we apply one tailed test.

Tabulated value of z at 5% level of significance for one tailed test is 1.645.

Since calculated value of z > tabulated value of z, we reject the null hypothesis.

i.e., The company's claim that only 5% of their product is defective is not acceptable.

4) A die was thrown 9000 times and of these 3220 yield a 3 or 4. Is this consistent with the hypothesis that the die was unbiased?

Solution

Given n = 9000

P=proportion of success of getting 3 or 4 in 9000 throws

$$= \frac{3220}{9000}$$

=0.3578

P= Population proportion of success

=P(getting a 3 or 4)

=P(getting 3) + p(getting 4)

$$= \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}$$

P=0.3333

$$\therefore Q = 1 - P = 0.6667$$

Null hypothesis: H0: The die is unbiased

Alternative hypothesis H1: $\mathrm{P} \neq \frac{1}{3}$ (Two tailed alternative)

Test statistic

$$z = \frac{p - P}{\sqrt{\frac{PQ}{n}}}$$

$$z = \frac{0.3578 - 0.3333}{\sqrt{\frac{(0.3333)(0.6667)}{9000}}} = 4.94$$

Since alternative hypothesis is two tailed alternative, we apply two tailed test

The tabulated value of z for two tailed test at 5% level of significance is 1.96.

Since calculated value of z > tabulated value of z , the null hypothesis is rejected.

i.e., the die is biased.

5) A random sample of 500 pineapples was takes from a large consignment and 65 were found to be bad. Find the percentage of bad pineapples in the consignment.

Solution

Given n= 500

P=proportion of bad pineapples in the sample= $\frac{65}{500} = 0.13$

q=1-p=0.87

We know that limits for population proportion P are given by $p \pm 3\sqrt{\frac{pq}{n}}$

$$= 0.13 \pm 3\sqrt{\frac{0.13(.87)}{500}}$$

$$= .13 \pm 0.045 = (0.175, 0.085)$$

$\therefore$ The percentage of bad pineapples in the consignment lies between 17.5%, 8.5%

6) A random sample of 500 apples was taken from a large consignment and 60 were found to be bad. Obtain the 98% confidence limits for the percentage number of bad apples in the consignment.

Solution:

Given n = 500

P= proportion of bad apples in the sample = $\frac{60}{500} = 0.12$

q=.88

We know that 98% confidence limits for population proportion are

$$p \pm 2.33\sqrt{\frac{pq}{n}}$$

$$= 0.12 \pm 2.33\sqrt{\frac{0.12 \times .88}{500}}$$

$$= 0.12 \pm 2.33(0.01453)$$

$$= (0.08615, 0.15385)$$

$\therefore$ 98% confidence limits for percentage of bad apples in the consignment are (8.61%, 15.38%).

Difference of Proportions

Suppose 2 large samples of sizes n_1 and n_2 are taken respectively from 2 different populations.

To test the significant difference between the sample proportions p_1 and p_2

$$z = \frac{p_1 - p_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad \text{where} \quad p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} \quad \text{and } q = 1-p$$

Examples:

1) Random samples of 400 men and 600 women were asked whether they would like to have a flyover near their residence. 200 men and 325 women were in favour of the proposal. Test the hypothesis that proportions of men and women in favour of the proposal are same at 5% level of significance?

Solution

p1; p2

Given sample sizes

n_1=400, n_2=600

Proportion of men = $P_1 = \frac{200}{400} = 0.5$

Proportion of women $P_2 = \frac{325}{600} = 0.541$

Null hypothesis H_0: Assume that there is no significant difference between the option ofmen and women as far as proposal of flyover is connected

i.e., $H_0 = P_1 = P_2$

Alternative Hypothesis: H_1: $p_1 \neq p_2$ (two tailed alternative)

$$z = \frac{p_1 - p_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where $p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$ and q = 1-p

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

$$p = \frac{400 \times \frac{200}{400} + 600 \times \frac{325}{600}}{400 + 600}$$

$$= \frac{525}{1000} = 0.525$$

q=1-p=1-.525=0.475

$$z = \frac{0.5 - 0.541}{\sqrt{0.525(.475)\left(\frac{1}{400} + \frac{1}{600}\right)}}$$

$$= \frac{-0.041}{0.032}$$

= -1.28

$$|z| = 1.28$$

The tabulated value of z for two tailed test at 5% level of significance is 1.96.

Since calculated value of z < tabulated value , we accept the null hypothesis at 5% level of significance.

i.e., there is no difference of option between men and women as far as proposal of flyover is concerned.

2) Before an increase in excise duty on tea, 800 persons out of a sample of 1000 persons were found to be tea drinkers. After an increase on duty, 800 people were tea drinkers in a sample of 1200 people. Using standard error of proportion, state whether there is a significant decrease in the consumption of tea after the increase in excise duty?

Solution

Given n_1=1000, n_2=1200

$P_1 = \frac{800}{1000} = 0.8$

$P_2 = \frac{800}{1200} = 0.667$

Null hypothesis H_0: Assume that there is no significant difference between the consumption of tea before and after the increase in excise duty.

i.e., $H_0 = P_1 = P_2$

Alternative Hypothesis: H_1: $p_1 > p_2$ (one-tailed alternative)

$$z = \frac{p_1 - p_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where $p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$ and q = 1-p

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

$$p = \frac{1000 \times \frac{800}{1000} + 1200 \times \frac{800}{1200}}{1000 + 1200}$$

$= 0.727$

q=1-p=1-0.727=0.273

$$z = \frac{0.8 - 0.667}{\sqrt{0.727(0.273)\left(\frac{1}{1000} + \frac{1}{1200}\right)}}$$

$$= \frac{0.133}{0.019} = 7$$

$|z| = 7$

The tabulated value of z for one tailed test at 5% level of significance is 1.645.

Since calculated value of z > tabulated value , we reject the null hypothesis at 5% level of significance.

That is, there is a difference in the consumption of tea before and after the increase in excise duty.

Note:

If we want to test the significance of the difference between p1 and p where

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

$$z = \frac{p_1 - p}{\sqrt{\frac{n_2 pq}{n_1(n_1 + n_2)}}}$$

3) In a random sample of 400 students of the university teaching department, it was found that 300 students failed in the examination. In another sample of 500 students of the affiliated colleges the number of failures in the same examination was found to be 300. Find out whether the proportion of failures in the university teaching departments significantly greater than the proportion of failures in the university teaching departments and the affiliated colleges taken together.

Solution

Given n_1=400, n_2=500

$$P_1 = \frac{300}{400} = 0.75$$

$$P_2 = \frac{300}{500} = 0.6$$

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

$$p = \frac{400(0.75) + 500(0.6)}{400 + 500}$$

=0.667

q=0.333

Null hypothesis: H_0: Assume that there is no significant difference between p_1 and P (i.e., p_1=P)

Alternative Hypothesis: p_1>P (one-tailed alternative)

Test statistic $$z = \frac{p_1 - p}{\sqrt{\frac{n_2 pq}{n_1(n_1 + n_2)}}}$$

$$z = \frac{0.75 - 0.667}{\sqrt{\frac{500 \times 0.667 \times 0.333}{400(400+500)}}}$$

=4.74

The table value of z for one tailed test at 5% level of significance is 1.645.

Since calculated value of z > table value , we reject the null hypothesis.

Therefore the proportion of failures in the affiliated colleges is greater than the proportion of failures in university departments and affiliated colleges taken together.

Note:

i) Suppose the population proportions P_1 and P_2 are given and $P_1 \neq P_2$. If we want to test the hypotheis that the difference P1-P2 in the population proportions is likely to be hidden in simple samples of sizes n1 and n2 from the two populations respectively then

$$z = \frac{(p_1 - p_2) - (P_1 - P_2)}{\sqrt{\frac{P_1 Q_1}{n_1} + \frac{P_2 Q_2}{n_2}}}$$

ii) If the sample proportions are not known then we use

$$z = \frac{|P_1 - P_2|}{\sqrt{\frac{P_1 Q_1}{n_1} + \frac{P_2 Q_2}{n_2}}}$$

4) A cigarette manufacturing firm claims that its brand A line of cigarettes outsells its brand B by 8%. (The meaning of this one is P_1-P_2=8%=0.08)If it is found that 42 out of a sample of 200 smokers prefer brand A and 18 out of another sample of 100 smokers prefer brand B.Test whether the 8% difference is a valid claim.

Solution

Given n_1=200, n_2=100

$p_1 = \frac{42}{200}$

$p_2 = \frac{18}{100}$

P_1-P_2=8%= $\frac{8}{100} = 0.08$

Null Hypothesis: Assume that 8% difference in the sale of two brands of cigarettes is valid claim.

i.e., H_0 : P_1-P_2=0.08

Alternative hypothesis: $P_1 - P_2 \neq 0.08$ (Two-tailed alternative)

$$z = \frac{(p_1 - p_2) - (P_1 - P_2)}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

P=0.2

Z=-1.02

$|z| = 1.02$

Since the alternative is two tailed alternative, we apply two tailed test.

The table value of z at 5% level of significance for two tailed test is 1.96

Since the calculated value of z(=1.02) < table value(=1.96), we accept the null hypothesis.

Hence, 8% difference in the sale of two brands of cigarettes is valid claim.

5) In two large populations, there are 30% and 25% respectively of fair haired people. Is this difference likely to be hidden in samples of 1200 and 900 respectively from the two populations.

Solution

Given n_1=1200, n_2=900

$P_1 = \frac{30}{100} = 0.3$

$P_2 = \frac{25}{100} = 0.25$

Q_1= 1- P_1 = 0.7

Q2= 1-P_2 = 0.75

Null hypothesis: Assume that sample proportions are equal

i.e., H_0:p_1=p_2

i.e., the difference in population proportion is likely to be hidden in sampling

Alternative Hypothesis: $p_1 \neq p_2$ (Two-tailed alternative)

$$z = \frac{|P_1 - P_2|}{\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}}$$

=2.55

The table value of z at 5% level of significance for two tailed test is 1.96

Since calculated value of z > table value, we reject the null hypothesis.

i.e., The sample proportions are not equal.

Exercises:

5) A machine produced 20 defective articles in a batch of 400. After over hauling it produced 10 defective in a batch of 300. Has the machine improved? (Null hypo: $p_1=p_2$; Alter. hypo:$p_1<p_2$)

Test of Significance for single mean

Suppose we want to test whether the given sample of size n has been drawn from a population with mean μ, we set up the null hypothesis that there is no difference between t $\bar{x}$ and μ where $\bar{x}$ is the sample mean.

The test statistic z $= \dfrac{\bar{x}-\mu}{\left(\dfrac{s}{\sqrt{n}}\right)}$ where s is the sample size.

If σ is given

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

Note:

The values $\bar{x} \pm 1.96\dfrac{\sigma}{\sqrt{n}}$ are called 95% confidence limits for the means of the populations corresponding to the given sample.

Similarly, $\bar{x} \pm 2.58\dfrac{\sigma}{\sqrt{n}}$ are called 99% confidence limits.

Examples:

1) A sample of 900 members has a mean of 3.4cms and S.D. 2.61cms. Is the sample drawn from a large population of mean 3.25cm and S.D. 2.61cms? If the population is normal and its mean is unknown. Find the 95% fiducial limits of true mean?

Solution

Given n = 900 ; $\mu = 3.25 \;\; \sigma - 2{,}61$

$\bar{x} = 3.4$; $s = 2.61$

Null hypothesis: H_0: Assume that the sample has been drawn from the population with mean $\mu = 3.25$.

Alternative Hypothesis: H_1: $\mu \neq 3.25$ (Two tailed alternative)

The test statistic $z = \dfrac{\bar{x} - \mu}{\sigma / \sqrt{n}}$

=1.724

The table value of z for two tailed test at 5% level of significance is 1.96

Since $|z| = 1.724$ <1.96, we accept the null hypothesis.

That is the sample has been drawn from the population with mean $\mu = 3.25$.

95% confidence limits are $\bar{x} \pm 1.96 \dfrac{\sigma}{\sqrt{n}}$

= $3.4 \pm 1.96 \dfrac{2.61}{\sqrt{900}}$

=3.57 and 3.2295

2) An insurance agent has claimed that the average age of policy holders who issue through him is less than the average for all agents which is 30.5 years. A random sample of 100 policy holders who had issued through him gave the following age distribution

Age	16-20	21-25	26-30	31-35	36-40
No. of persons	12	22	20	30	16

Calculate the A.M. & S.D. of this distribution and use these values to test his claim at 5% level of significance.

Solution

Calculate mean and s.d for the above data we get

$\bar{x} = 28.8$, S.D. s = 6.35

Null hypothesis H_0: The sample is drawn from a population with mean $\mu = 30.5$

Alternative hypothesis: H_1: $\mu < 30.5$ (see the question there less than came) (one tailed)

$$z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

$$= \frac{28.8 - 30.5}{\frac{6.35}{\sqrt{100}}}$$

= -2.677

Since the alternative hypothesis is one tailed alternative we apply one tailed test.

The table value of z at 5% level of significance for one tailed test is 1.645

Since calculated value $|z| = 2.68 >$ table value 1.645, we reject the null hypothesis.

$\therefore \bar{x}$ and μ differ significantly.

i.e., the sample is not drawn from a population with mean μ.

IV:Test of significance for difference of Means

Let $\overline{x_1}$ be the mean of a sample of size n_1 from a population with mean μ_1 and S.D. ${\sigma_1}^2$

Let $\overline{x_2}$ be the mean of a sample of size n_2 from a population with mean μ_2 and S.D. ${\sigma_2}^2$

To test whether there is any significant difference between $\overline{x_1}$ and $\overline{x_2}$.

$$z = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}}.$$

Note: If the samples have been drawn from the same population then

$${\sigma_1}^2 = {\sigma_2}^2 = \sigma^2$$

$$z = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}}$$

ii) If σ is not known then

$$\sigma^2 = \frac{n_1{\sigma_1}^2 + n_2{\sigma_2}^2}{n_1 + n_2}$$

Examples:

1) The means of 2 large samples 1000 and 2000 members are 67.5 inches and 68.0 inches respectively. Can the samples be regarded as drawn from the population of S.D. 2.5 inches?

Solution

Given n_1=1000; n_2=2000

$\overline{x_1} == 67.5; \overline{x_2} = 68$

Population S.D. $\sigma = 2.5$ inches

Null hypothesis H_0: $\overline{x_1} = \overline{x_2}$.

Alternative hypothesis H_1: $\overline{x_1} \neq \overline{x_2}$ (two tailed alternative)

$$z = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}}$$

$$z = \frac{67.5 - 68}{\sqrt{\frac{(2.5)^2}{1000} + \frac{(2.5)^2}{2000}}} = \frac{-0.5}{0.0968}$$

=-5.16

Since the alternative hypothesis is two tailed alternative, we apply two-tailed test.

The table value of z for two tailed test at 5% level of significance is 1.96.

$|z| = 5.16 > 1.96$

$\therefore$ we reject the null hypothesis at 5% level of significance .

i.e., the samples are not drawn from the same population of S.D. 2.5 inches.

2) The mean yield of wheat from a district was 210 pounds with S.D. 10 pounds per acre from a sample of 100 plots. In another district the mean yield was 220 pounds with S.D. 12 pounds from a sample of 150 plots. Assuming that the S.D. of yield in the entire state was 11 pounds, test whether there is any significant difference between the mean yield of crops in the two districts.

Solution

$\overline{x_1} = 210$ n_1 =100

$\overline{x_2} = 220$; n_2 = 150

$\sigma = 11$

Null hypothesis H_0: $\overline{x_1} = \overline{x_2}$

Alternative hypothesis H_1: $\overline{x_1} \neq \overline{x_2}$

$$z = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}}$$

$$= \frac{210 - 220}{\sqrt{\frac{11^2}{100} + \frac{11^2}{150}}}$$

=7.041

Since $|z| = 7.041 > 1.96$ at 5% level of significance, we reject the null hypothesis.

That is there is a significant difference between the mean yield of crops in the two districts.

Note: If the two samples are drawn from two populations with unknown standard deviations then

$$z = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

3) In a survey of buying habits, 400 women shoppers are chosen at random in super market A located in a certain section of the city. Their average weakly food expenditure is Rs.250 with a S.D. of Rs.40. For 400 women shoppers chosen at random in super market B in another section of the city, the average weakly food expenditure is Rs.220 with a S.D. of Rs.55. Test at 1% level of significance whether the average weakly food expenditure of the two populations of shopper are equal.

Solution

n_1=400; $\overline{x_1} = 250$, $s_1 = 40$

n_2=400; $\overline{x_2} = 220$, $s_2 = 55$

Null hypothesis H_0: $\overline{x_1} = \overline{x_2}$

Alternative hypothesis H_1: $\overline{x_1} \neq \overline{x_2}$

$$z = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$z = \frac{250 - 200}{\sqrt{\frac{40^2}{400} + \frac{55^2}{400}}}$$

z= 8.82

The table value of z for two tailed test at 1% level of significance is 2.58.

Since $|z| = 8.82 > 2{,}58$, we reject the null hypothesis.

That is, The average weakly food expenditure of the two populations of shoppers are not equal.

4) The means of two samples of 1000 and 2000 items are 67.5 and 68.0 respectively. Can the samples be regarded at 5 % level of significance, as drawn from the same population with standard deviation 2.5?

Solution

$n_1 = 1000$, $n_2 = 2000$

$\overline{x_1} = 67.5; \overline{x_2} = 68$

$\sigma = 2.5$

Null hypothesis H_0 : $\overline{x_1} = \overline{x_2}$

Alternative hypothesis H_1: $\overline{x_1} \neq \overline{x_2}$

$$z = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}}$$

=5.163

Since $|z| > 1.96$, we reject the null hypothesis.

Therefore The samples are not drawn from the same population with standard deviation 2.5

Exercises:

5) A sample of 100 electric light bulbs produced by manufacturer A showed a mean life time of 1190 hrs and a S.D of 90 hrs. A sample of 75 bulbs produced by manufacturer B showed a mean life time of 1230 hrs with a standard deviation of 120 hrs. Is there a difference between the mean life time on the two brands at significance level of i) 5% , ii) 1% (Ans: z= 2.421) (In this problem Samples are draw from different population)

6) There are two brands of car tyres A and B in the market. A sample of 100 tyres of brand A has an average life of 37500 kms with a S.D. of 2500 kms. Another sample of 75 tyres of brand B has an average life of 3900 kms with a S.D. of 3000 km. Can we conclude that brand B is better than better than A? (In this problem samples are drawn from same population)

In alternative hypothesis

Brand B mean value is $\overline{x_2}$ Brand A mean value is $\overline{x_1}$

Given brand B is better than better than A so we write $\overline{x_2} > \overline{x_1}$

i.e., $\overline{x_1} < \overline{x_2}$

UNIT-V

Definition: A sample of size less than or equal to 30 is said to small sample.

Testing the significance of the difference of sample mean

t-test

$t = \dfrac{\bar{x} - \mu}{\dfrac{s}{\sqrt{n-1}}}$ where $\bar{x} = \dfrac{\sum x}{n}$; $s^2 = \dfrac{\sum \left(x - \bar{x}\right)^2}{n}$

The number of degree of freedom of statistic generally denoted by γ is defined as the number n of independent observations in the sample (i.e., the sample size) minus the no. k, of population parameters which must be estimated from sample observations.

$\gamma = n - k$

Degree of freedom = n-1

Assumptions:

1) The parent population from which the samples are drawn are normally distributed.

2) The two samples are random and independent of the other.

3) ${\sigma_1}^2 = {\sigma_2}^2 = \sigma^2$ i.e., the population variances are equal.

Examples:

A sample of 26 bulbs gives a mean life of 990 hours with a standard deviation of 20 hours. The manufacturer claims that the mean life of bulbs is 1000 hours. Is the sample not upto the standard

Solution

$\bar{x} = 990, \mu = 1000, \mathrm{s} = 20$

Null hypothesis: The sample is upto the standard.

Alternative hypothesis: $\mu < 1000$

$t = \dfrac{\bar{x} - \mu}{\dfrac{s}{\sqrt{n-1}}}$

=-2.5

$|t| = 2.5$

The tabulated value of t at 5% level of significance for 25 degree of freedom of one –tailed test is 1.7

Degree of freedom = n-1 =26-1=25

Since the calculated value of t is greater than tabulated value of t we reject the null hypothesis H0.

Therefore the sample is not upto the standard.

2) Tests made on the breaking strength of hard drawn copper wire gave the following results in kilograms

58, 582, 580, 578, 582, 580, 582, 606, 594

It was stated that the population mean of the breaking strength is 593kgs. Examine the validity of this statement given that the 5% value of the student's t test for degree of freedom, 9 and 10 are respectively 2.31, 2.26, 2.23.

Solution

Null hypothesis: H_0: $\mu = 593$

Alternative hypothesis H_1: $\mu \neq 593$

Calculation of mean $\bar{x}$ and standard deviation s

	$x - \bar{x}$	$(x - \bar{x})^2$
588	2.222	4.937
582	-3.778	14.273
580	-5.778	33.385
578	-7.778	60.497
582	-3.778	14.273
580	-5.778	33.385
582	-3.778	14.273
606	20.222	408.929
594	8.222	67.601
Total		651.553

$$\bar{x} = \frac{\sum x}{n} = 585.778$$

$$s^2 = \frac{\sum (x - \bar{x})^2}{n} = \frac{651.553}{9} = 72.395$$

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n}} = 8.509$$

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n-1}}} = \frac{585.778 - 593}{\left(\frac{8.509}{2.8284}\right)} = \frac{-7.222}{3.0084}$$

=2.4006

Degree of freedom = 9-1=8

The table value of t for 8 degree of freedom at 5% level of significance is 2.31.

Since $|t|$ =2.4006 > 2.31, we reject the null hypothesis

Therefore the population mean of the breaking strength is not 593 kgs

Testing the significance of the difference between two sample means

$$t = \frac{\overline{x_1} - \overline{x_2}}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Degree of freedom = n_1+n_2-2

Examples:

1) The I.Q's of 16 students from an area of the city showed a mean of 107 with S.D. of 10 while the I.Q. of 14 students from another area of the city showed a mean of 112 and S.D. of 8. Is there a significant difference between the I.Q.'s of 2 groups at 5% level of significance?

Solution

n_1= 16, n_2=14

$\overline{x_1} = 107$ $\overline{x_2} = 107$

s_1=10, s_2=8

Null hypothesis H_0: $\mu_1 = \mu_2$

There is no significant difference between the sample means.

Alternative hypothesis H_1: $\mu_1 \neq \mu_2$

Test statistic

$$t = \frac{\overline{x_1} - \overline{x_2}}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad s^2 = \frac{n_1 {s_1}^2 + n_2 {s_2}^2}{n_1 + n_2 - 2}$$ where

$$s^2 = \frac{n_1 {s_1}^2 + n_2 {s_2}^2}{n_1 + n_2 - 2}$$ =9.44

t=-1.447

$|t| = 1.447$

Degree of freedom = $n_1+n_2-2=16+14-2=28$

The table value of t for 2 degree of freedom is 2.05

The calculated value of t is less than the table value of t 2.05.

Hence the null hypothesis is accepted.

There is no significant difference between two I.Q.'s.

2) A group of 10 boys fed on diet A and another group of boys fed on diet B recorded the following increase in weights

Diet A	5	6		1	12	4	3	9	6	10kgs
Diet B	2	3	6	8	10	1	2	8kgs		

Does it show the superiority of diet A over that of diet B?

Solution

Let the increase in weights (in kgs) due to diets A and B be denoted by the variables x and y respectively.

Null hypothesis: H_0: $\mu_1 = \mu_2$

Alternative hypothesis H_1: $\mu_1 > \mu_2$

Computations of the sample mean and Standard deviation

X_1	$x_1 - \overline{x_1}$	$\left(x_1 - \overline{x_1}\right)^2$	X_2	$x_2 - \overline{x_2}$	$\left(x_2 - \overline{x_2}\right)^2$
5	-1.4	1.96	2	-3	9
6	-0.4	0.16	3	-2	4
8	1.6	2.56	6	1	1
1	-5.4	29.16	8	3	9
12	5.6	31.36	10	5	25
4	-2.4	5.76	1	-4	16
3	-3.4	11.56	2	-3	9
9	2.6	6.76	8	3	9
6	-0.4	0.16			
10	3.6	12.96			
Total:64		102.4	40		82

$$\overline{x_1} = \frac{\sum x_1}{n_1} = \frac{64}{10} = 6.4$$

$$\overline{x_2} = \frac{\sum x_2}{n_2} = 5$$

$$s_1^2 = \frac{\sum (x_1 - \overline{x_1})^2}{n_1} = \frac{102.4}{10} = 10.24$$

$$s_2^2 = \frac{\sum (x_2 - \overline{x_2})^2}{n_2} = 10.25$$

$$s^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2} = \frac{102.4 + 82}{16} = 11.525$$

s=3.395

$$t = \frac{\overline{x_1} - \overline{x_2}}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

=0.8694

Degree of freedom = $n_1+n_2-2=16$

The table value of t for 16 degree of freedom is 1.75

Since the calculated value of t is less than the tabulated value, it is not significant.

Hence the null hypothesis is accepted.

We conclude that the diets A and B do not differ significantly in respect of increase in weights.

Testing the significance of difference of sample variances

F – test

Let n_1 be the number of observations in a sample I from the first population with variance σ_1^2 and n_2 be the number of observations in a sample II from the second population with variance σ_2^2.

We set up null hypothesis H_o: $\sigma_1^2 = \sigma_2^2 = \sigma^2$

$$F = \frac{S_1^2}{S_2^2}$$

Where

$$S_1^2 = \frac{\sum (X - \overline{X_1})^2}{n_1 - 1}$$

$$S_2{}^2 = \frac{\sum\left(X - \overline{X_2}\right)^2}{n_2 - 1}$$

Degree of freedom = (n_1-1, n_2-1)

Note:

In numerical problems, we take the greater of the variances $S_1{}^2$ or $S_2{}^2$ in the numerator.

Assumptions on F-test

1) The populations for each sample must be normally distributed

2) The samples must be random and independent.

3) The ratio of $\sigma_1{}^2$ to $\sigma_2{}^2$ should be equal to 1 or greater than 1.

Examples:

1) Two samples of 6 and 7 items have the following values of the variables

Sample I	39	41	43	41	45	39	
Sample II	40	42	40	44	39	38	40

Do the sample variances vary significantly?

Solution

Null hypothesis H_0: $\sigma_1{}^2 = \sigma_2{}^2$

The sample variances do not differ significantly

Alternative hypothesis H_1: $\sigma_1{}^2 \neq \sigma_2{}^2$

$$F = \frac{S_1{}^2}{S_2{}^2}$$

X_1	$x_1 - \overline{x_1}$	$\left(x_1 - \overline{x_1}\right)^2$	X_2	$x_2 - \overline{x_2}$	$\left(x_2 - \overline{x_2}\right)^2$
39	-2.3	5.29	40		0.16
41		0.09	42	1.6	2.56
43	1.7	2.89	40	-0.4	0.16
41	-0.3	0.09	44	3.6	12.96
45	3.7	13.69	39	-1.4	1.96
39	-2.3	5.29	38	-2.4	5.76
			40	-0.4	0.16
Total		27.34			23.72

$$\overline{x_1} = \frac{\sum x_1}{n_1} = \frac{248}{6} = 41.3$$

$$\overline{x_2} = \frac{\sum x_2}{n_2} = \frac{283}{7} = 40.4$$

$$S_1{}^2 = \frac{\sum \left(X - \overline{X_1}\right)^2}{n_1 - 1}$$

=5.468

$$S_2{}^2 = \frac{\sum \left(X - \overline{X_2}\right)^2}{n_2 - 1}$$

=3.9533

$$F = \frac{S_1{}^2}{S_2{}^2}$$

=1.383

Degree of freedom = (n_1-1, n_2-1)=(5,6)

The table value of F for (5,6) degree of freedom at 5%level of significance is 4.39

Since calculated value is less than the table value, we accept the null hypothesis.

Hence the variances do not differ significantly.

2) Two independent samples of sizes 7 and 6 have the following values of the variables

Sample I	28	30	32	33	31	29	34
Sample II	29	30	30	24	27	28	-

Examine whether the samples have been drawn from normal population having the same variance?

Solution:

F=S2^2/S1^2

F=1.113

The table value of F for the degree of freedom (5,6) at 5%level of significance is 4.39

Since the calculated value of F is less than the table value, we accept the null hypothesis.

Hence, the samples have been drawn from normal population having the same variance.

3) Two random samples gave the following results.

Sample	Size	Sample mean	Sum of squares of deviations from the mean $\sum(x-\bar{x})^2$
I	10	15	90
II	12	14	108

Test whether the samples could have come from the same normal population

Soln:

Null hypothesis:

Alternative hypothesis:

$S1^2= \dfrac{\sum\left(x-\overline{x_1}\right)^2}{n_1-1}=\dfrac{90}{9}=10$

$S_2{}^2=\dfrac{\sum\left(x_2-\overline{x_2}\right)^2}{n_2-1}=\dfrac{108}{11}=9.8$

F=S1^2/S2^2=10/9.8=1.02

Degree of freedom is (9,11)

The table value of F for the degree of freedom (9,11) is 2.90

Since calculated value is less than the table value, we accept the null hypothesis.

Hence the samples could have come from the same normal population.

Chi-Square Test-

Type-I: Goodness of fit

1) The following table gives the number of accidents that take place in an industry during various days of the week. Test if accidents are uniformly distributed over the week

Day	Mon	Tue	Wed	Thur	Fri	Sat
No. of accidents	14	`18	12	11	15	14
Degree of freedom	5	6	7			
Value of χ^2	11.07	12.59	14.07			

Soln: Null hypotheis: The accidents are uniformly distributed over the week

Alternative hypothesis: The accidents are not uniformly distributed over the week

Total no. of accidents=84

We have to test the hypothesis that the accidents are uniformly distributed over the 6 days of the week.

On the basis of this hypothesis we should expect 84/6=14 accidents each day.

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

O	E	$(O-E)^2$	$(O-E)^2/E$
14	14	0	0
18	14	16	1.142
12	14	4	4/14=0.285
11	14	9	9/14=0.642
15	14	1	1/14=0.071
14	14	0	0
		Total	2.14

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 2.14$$

Degree of freedom=n-1=6-1=5

The table value of chi-square for the degree of freedom 5 at 5% level of significance is 11.07.

Since the calculated value of χ^2 is less than the table value we accept the null hypothesis.

Hence the accidents are uniformly distributed over the week.

2) In a cross breeding experiment with plants of certain species, 240 offspring were classified into 4 classes with respect to the structure of their leaves as follows:

Class:	I	II	III	IV	Total
Frequency	21	127	40	52	240

According to the theory of heredity, the probability of the four classes should be in the ratio 1:9:3:3. Are these data consistent with theory.

Soln:

Null hypothesis: The given data are consistent with the theory that the frequencies in the four classes should be in the ratio 1:9:3:3.

On the basis of this hypothesis , the expected frequencies are

$$\frac{1}{16}\times 240, \frac{9}{16}\times 240, \frac{3}{16}\times 240, \frac{3}{16}\times 240$$

i.e., 15,135,45,45

Computation of χ^2 :

O	E	$(O-E)^2$	$(O-E)^2/E$
21	15	36	2.4
127	135	64	0.4742
40	45	25	0.5556
52	45	49	1.089
			4.5188

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 4.5188$$

Degree of freedom=n-1=4-1=3

The table value of χ^2 for the degree of freedom 3 at 5% level of significance is 7.81

Since the calculated value of χ^2 is less than the table value we accept the null hypothesis.

Hence the given data are consistent with the theory that the frequencies in the four classes should be in the ratio 1:9:3:3.

3) Records taken of the number of male and female births in 800 families having four children are as follows:

No. of boys	0	1	2	3	4
No. of girls	4	3	2	1	0
No. of families	32	178	290	236	64

Test whether the data are consistent with the hypothesis that the binomial law holds and that the chance of male birth is equal to that of a female birth, namely p= q =1/2

Soln:

Null hypothesis: The data is consistent with the binomial law of equal probability for male and female births.

i.e., p=q=1/2

Calculation of theoretical frequencies:

P(r)= Prob. of r male births in a family of n births

$$= {}^{n}c_r\, p^r q^{n-r} = {}^{n}c_r \left(\frac{1}{2}\right)^r \left(\frac{1}{2}\right)^{n-r}$$

$$p(r) = n_{c_r}\left(\frac{1}{2}\right)^n$$

Frequency of r male births f(r)=N p(r)= $8004_{c_r}\left(\frac{1}{2}\right)^4 = 50\times 4_{c_r}$

Put r=0,1,2,3,4 we get

f(0)= $50\times 4_{c_0} = 50$

f(1)= $50\times 4_{c_1} = 200$

f(2)= $50\times 4_{c_2} = 300$

f(3)= $50\times 4_{c_3} = 200$

f(4)= $50\times 4_{c_4} = 50$

Calculation of χ^2 :

O	E	$(O-E)^2$	$(O-E)^2/E$
32	50	324	6.48
178	200	484	2.42
290	300	10	0.3333
236	200	1296	6.48
64	50	196	3.92
			19.63

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 19.63$$

Degree of freedom=5-1=4

The table value of χ^2 for the degree of freedom 4 at 5% level of significance is 9.49

Since the calculated value of χ^2 is greater than the table value we reject the null hypothesis.

Hence the data are not consistent with the hypothesis that the binomial law holds and that the chance of male birth is not equal to that of a female birth.

4) In the accounting department of a bank, 100 accounts are selected at random and examined for errors. The following result has been obtained.

No. of errors:	0	1	2	3	4	5	6
No. of accounts:	35	40	19	2	0	2	2

Does the information verify that the errors are distributed according to the poisson distribution law?

Soln:

$p(r) = \frac{e^{-m} m^r}{r!}$, m-mean

To get mean, use $\frac{\sum fr}{\sum f}$

If the random variable X denotes the number of errors, the given distribution is as follows:

x	f	fx
0	35	0
1	40	40
2	19	38
3	2	6
4	0	0
5	2	10
6	2	12

$$\text{Mean m} = \frac{\sum \text{fx}}{\sum \text{f}} = \frac{106}{100} = 1.06$$

m=1.06

To fit a poisson distribution to the data we take the parameter of poisson distribution is equal to the mean.

i.e., m=1.06

The frequency of x errors is given by the poisson law as

f(x)=Np(x)= $N \frac{e^{-m} m^x}{x!}$

f(0)= $100 \frac{e^{-1.06} (1.06)^0}{0!} = 34.65$

f(1)= $100 \frac{e^{-1.06} (1.06)^1}{1!} = 36.73$

f(2)=19.47; f(3)=6.88, f(4)=1.82, f(5)=.39, f(6)=0.06

Calculation of χ^2 :

O		E		$(O-E)^2$	$(O-E)^2/E$
35		34.65		.1225	0.0035
40		36.73		10.6929	.2911
19		19.47		0.2209	0.0113
2	6	6.88	9.15	$(6-9.15)^2=9.9225$	1.0844
0		1.82			
2		.39			
2		0.06			
					1.3903

(In observed frequency Any frequency value is less than 5 we make the adjustment with other classes)

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 1.3903$$

Degree of freedom=n-1-(here we calculated mean value) so we have

=n-1-1

Again 3 cell are reduced. So degree of freedom = n-1-(1)-3=7-1-(1)-3=2

The table value of χ^2 for 2 degree of freedom at 5% level of significance is 5.99.

Since the calculated value of χ^2 is less than the table value we accept the null hypothesis.

Hence the information verify that the errors are distributed according to the poisson distribution.

Chi-square- as a test of independence

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Expected frequency= $\dfrac{\text{RowTotal} \times \text{ColumnTotal}}{\text{GrandTotal}}$

The degree of freedom = (m-1)(n-1) where m is no of rows and n is no. of columns

Remark: For 2 x 2 contingency table

a	b
c	d

$$\chi^2 = \frac{N(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}$$

Yates correction:

Any cell frequency value is less than 5 we apply yates correction.

Add 0.5 to that cell and adjust the remaining cell frequencies.

Examples:

1) The following table gives the no. of good and bad parts produced by each of the three shifts in a factory.

	Good parts	Bad parts	Total
Day shift	960	40	1000
Eve shift	940	50	990
Night shift	950	45	995
	2850	135	2985 (2985)

Test whether or not the production of bad parts is independent of the shift on which they were produced?

Chi-square value for 2 degree of freedom is 5.991

Soln:

Null hypothesis: The production of bad parts is independent of the shift

Alternative hypo: The production of bad parts is not independent of the shift.

Now we calculate Expected frequencies:

$$\text{Expected Frequency} = \frac{\text{RowTotal} \times \text{ColumnTotal}}{\text{GrandTotal}}$$

E(960)= $\frac{1000 \times 2850}{2985} = 954.77$

E(40)= $\frac{1000 \times 2850}{2985} = 45.23$

E(940)= $\frac{990 \times 2850}{2985} = 945.23$

E(50)= $\frac{990 \times 135}{2985} = 44.77$

E(950)= $\frac{995 \times 2850}{2985} = 950$

E(45)=45

Calculation of χ^2

Observed Frequency O	Expected Frequency E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
960	954.77	27.35	0.0286
40	45.23	27.35	0.6047
940	945.23	27.35	0.0289
50	44.77	27.35	0.6109
950	950	0	0
45	45	0	0
			1.2731

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

=1.2731

The degree of freedom = (m-1)(n-1) where m is no of rows and n is no. of columns

=(3-1)(2-1)=2

The table value of chi-square for 2 degree of freedom is 5.991

Since calculated value is less than table value, we accept the null hypothesis.

Hence the production of bad parts is independent of the shift on which they were produced.

2) Can vaccination be regarded as preventive measure of small pox as evidenced by the following data:

Of 1482 persons exposed to small pox in a locality, 368 in all were affected. Of these 1482 persons, 343 were vaccinated and of these only 35 were attacked. Given chi-square at 5% level of significance for 1 degree of freedom is 3.841.

Soln:

	Vaccinated	Not Vaccinated	Total
Attacked by small pox	35	(368-35)=333	368
Not attacked	(343-35)=308	? (1114-308)=806	?(1482-368)=1114
	343	333+806=1139	1482

Null hypothesis: vaccination and attack by small pox are independent.

Calculation of Expected Frequencies:

$$\text{Expected Frequency} = \frac{\text{RowTotal} \times \text{ColumnTotal}}{\text{GrandTotal}}$$

$$\text{E(35)} = \frac{368 \times 343}{1482} = 85.17$$

$$\text{E(333)} = \frac{368 \times 1139}{1482} = 282.83$$

$$\text{E(308)} = \frac{1114 \times 343}{1482} = 257.83$$

$$\text{E(806)} = \frac{1114 \times 1139}{1482} = 856.17$$

Calculation of χ^2

O	E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
35	85.17	2517.03	29.55
333	282.83	2517.03	8.90
308	257.83	2517.03	9.76
806	856.17	2517.03	2.94
		Total:	51.15

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 51.15$$

The degree of freedom for chi-square test = (m-1)(n-1)= (2-1)(2-1)=1

The table value of chi-square test for 1 degree of freedom is 3.841.

Since the calculated value 51.15> table value, we reject the null hyposthesis.

We conclude that vaccination and attacked by small pox are not independent.

Yates correction:

Any cell frequency value is less than 5 we apply yates correction.

Add 0.5 to that cell and adjust the remaining cell frequencies.

3) In an experiment on the immunization of goats from anthrax. The following results were obtained. Derive your inference on the efficiency of the vaccine.

	Died	Survived	Total
Inocculated	2 <5	10	12
Not inoculated	6	6	12
	8	16	24

Solution:

Since the cell (1,1) 2 which has the frequency 2 less than 5 we apply yates correction.

We add 0.5 to that cell and adjust the remaining cell frequencies.

After applying yate's correction

	Died	Survived	Total
Inocculated	2 .5 here we added 0.5	10 here we do subtraction with 0.5 is changed to 9.5	12 to get the same total
Not inoculated	6 now this one changed to 5.5	6 is increased to 0..5 Now this chaged to 6.5	12
	8	16	24

$$\text{ExpectedFrequency} = \frac{RowTotal \times ColumnTotal}{GrandTotal}$$

E(2.5)= $\frac{10 \times 8}{24} = 3.33$

E(9.5)= $\frac{12 \times 16}{24} = 8$

E(6.5)= $\frac{12 \times 8}{24} = 4$

E(5.5)= $\frac{12 \times 16}{24} = 8$

Calculation of χ^2

O	E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
2.5	3.33	0.68	0.204
9.5	8	2.25	0.28
6.5	4	6.25	1.56
5.5	8	6.25	0.78
			1.7

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 1.7$$

The degree of freedom for chi-square test = (m-1)(n-1)= (2-1)(2-1)=1

The table value of chi-square test for 1 degree of freedom is 3.841.

Since the calculated value 1.7 < table value, we reject the null hypothesis.

We conclude that inoculation is not effective against the disease.