M.Sc.

Information Technology

(DISTANCE MODE)

DIT 111

Probability and Queuing Theory

I SEMESTER

COURSE MATERIAL

Centre for Distance Education

Anna University Chennai

Chennai – 600 025

Author

Mrs. Maya Joby
Lecturer
Department of Mathematics
DMI College of Engineering
Chennai – 602 103

Reviewer

Dr.M.Srinivasan
Professor & Controller of Examinations
SSN College of Engineering,
Old Mahabalipuram Road,
Kalavakkam
Chennai. - 603110

Editorial Board

Dr. P. MANNAR JAWAHAR, Ph.D., M.E., DMIT. B.Sc.,
MISTE. MISME, MISNDT, MSAE(USA). FIE, Chartered Engineer

VICE-CHANCELLOR

ANNA UNIVERSITY CHENNAI
Chennai - 600 025, INDIA

September, 2008

FOREWORD

The prime objective of education is conversion of human being into intellectual capital of a nation. Conventional education system has its own limitations to achieve this objective. Realizing this fact all over the world more and more traditional universities are rapidly transforming themselves from single mode of traditional education to dual mode of both traditional as well as distance education. Distance Education is becoming an accepted and indispensable part of the main stream of the educational system of any nation. Technology has made it possible to provide the best and the most up-to-date education at a reasonable cost and without geographical boundaries.

Anna University Chennai has started the Centre for Distance Education during 2006 with the aim of providing equity access to quality professional education to all the deserving aspirants. The Centre for Distance Education, Anna University Chennai has introduced MBA in General Management, Technology Management, Retail Management, Human Resource Management, Financial Services Management and Health Services Management, MCA with emphasis on Banking Technology, Call Centre Management, E-Learning, Trading and Equity Management, Health Care Management, and M.Sc in Computer Science, Information Technology, Software Engineering and Computer Technology. These programmes have been well received by all the stake holders. Anna University Chennai has absorbed large number of learners from multiple segments for all the programmes introduced. Yes, we have achieved impressive success to celebrate, and that induce us to perform much better, in the days to come.

The course materials play vital role in imparting knowledge, specially in distance education. These materials delivered to you were prepared by experts in the respective areas. Authors have prepared the course materials based on learner centric approach, fine tuned by self instructional mode. I am sure these materials will meet your learning requirements in full by inculcating a new learning culture.

As we move towards more knowledge intensive economy, acquiring and sustaining relevant skills and knowledge is becoming increasingly significant.

On this line Anna University Chennai, shall continue to contribute its best and thereby enable our nation a much more knowledge rich nation.

My hearty congratulations and best wishes to all.

16/9/08

(P. MANNAR JAWAHAR)

Sardar Patel Road, Guindy, Chennai - 600 025, Tamilnadu, INDIA
Phone:+ 91-44-2235 1445, 2235 3445 Fax: +91-44-2235 0397 e-mail: vc@annauniv.edu

ACKNOWLEDGEMENT

I, Ms. Maya Joby, express my sincere thanks and deep sense of gratitude to the following persons who have constantly encouraged and supported me in this wonderful Endeavour.

Dr. Chellappan – Professor. Department of Computer Science and Engineering.
Deputy Director, Centre for Distance Education Anna University
For guiding and encouraging me during this course material preparation.

Dr. M.Gopal – Principal DMI College of Engineering, Chennai
for the institutional support rendered and encouraging me during the preparation of this course material.

All staff and authorities of Centre for Distance Education Anna University for providing me this opportunity to prepare this course material.

While preparing this material I have benefited immensely by referring to many books. I express my gratitude to all such authors and publishers.

Reference Books:

1). T. Veerarajan, “Probability, Statistics and Random Process” Tata McGraw Hill.
2) P. Kandasamy, K. Thilagavathi and K Gunavathi “ Probability Random variables and Random processors”. S Chand.
3) S.C. Gupta and V K Kapoor, “Fundamentals of Mathematical Statistics” Sultan Chand & Sons.

I wish to acknowledge my sincere thanks to my colleagues in DMI college of Engineering for guiding and helping me whenever I was stuck in trying to better explain some topics.

I also wish to thank my family members for their moral support and encouragement while preparing this material.

Mrs. Maya Joby
Lecturer
Department of Mathematics
DMI College of Engineering
Chennai – 602 103

DIT 111 PROBABILITY AND QUEUING THEORY

UNIT I

Probability and Random Variables: Probability concepts – Random variables – Moment generating function – Binomial, Poisson, Geometric, Uniform exponential, Normal distributions – Functions of Random variables.

UNIT II

Two-Dimensional Random Variables: Marginal and conditional distributions – Covariance – correlation and Regression – Transformation of Random Variables.

UNIT III

Tests of Hypothesis: Sampling distributions – Tests based on Normal, t and F distributions for means, variance and proportions, chi-square test for variance, independence and goodness of fit.

UNIT IV

Random Process: Classification – stationary process – Markov process – Poisson process – Markov chains.

UNIT V

Queueing Theory: Single and multi-server Markovian Queues – Stationary for queue size distributions – Little's formula – Average measures.

TEXT BOOK

1. T.Veerarajan, "Probability, Statistics and Random Process, Tata McGraw Hill, 2002.

REFERENCES

1. Taha, H.A. "Operations Research : An Introduction", Prentice Hall, New Delhi, 2002.
2. P. Kandasamy, K. Thilagavathi and K. Gunavathi, "Probability, Random Variables and Random Processors", S. Chand, 2003.

DIT 111 PROBABILITY AND QUEUEING THEORY

NOTES

UNIT 1

PROBABILITY AND RANDOM VARIABLES

- Introduction
- Probability Concepts
- Random Variables
- Expectation and Moments
- Moment Generating Functions
- Binomial Distribution
- Poisson Distribution
- Geometric Distribution
- Uniform Distribution
- Exponential Distribution
- Normal Distribution
- Functions of Random Variable

1.1 INTRODUCTION

Probability theory is an important part of contemporary mathematics. It plays a key role in the insurance industry, in the modeling of financial markets, and in statistics generally — including all those fields of endeavor to which statistics is applied (e.g. health, physical sciences, engineering, economics). The 20th century has been an important period for the subject, because we have witnessed the development of a solid mathematical basis for the study of probability, especially from the Russian school of probability under the leadership of A N Kolmogorov. We have also seen many new applications of probability — from applications of stochastic calculus in the financial industry to Internet gambling. At the beginning of the 21st century, the subject offers plenty of scope for theoretical developments, modern applications and computational problems. There is something for everyone in probability!

NOTES

1.2 LEARNING OBJECTIVES

The students will acquire

- Knowledge to define experiment, outcome, event, probability and equally likely.
- State the formula for finding the probability of an event.
- Knowledge to evaluate outcomes and probabilities for several simple experiments.
- Knowledge to Recognize the difference between outcomes that are, and are not, equally likely to occur.
- Knowledge to apply basic probability principles to solve problems.
- Knowledge to analyze each problem to identify the given information.
- Knowledge to identify the concepts and procedures needed to solve each problem.
- Knowledge to apply probability concepts to solve complex problems.
- Knowledge to identify connections between probability and the real world.
- Develop strong problem-solving skills.
- Familiarity with some of the distributions commonly used to represent real-life situations.

First of all let us go through some basic concepts in probability. Try to be thorough in the basics so you may feel comfortable with the subject.

1.3 PROBABILITY CONCEPTS

1.3.1 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 one of several outcomes. Thus an action which can produce any result or outcome is called a RANDOM EXPERIMENT. When a random experiment is performed each time it is called a TRIAL and the outcomes are known as EVENTS or CASES.

An experiment whose outcome or result can be predicted is called a Deterministic Experiment.

Sample Space : The sample space is an exhaustive list of all the possible outcomes of an experiment. Each possible result of such a study is represented by one and only one point in the sample space, which is usually denoted by S.

Examples

Experiment of Rolling a die once:

NOTES

Sample space S = {1,2,3,4,5,6}

Experiment Tossing a coin:

Sample space S = {Heads,Tails}

Experiment Measuring the height (cms) of a girl on her first day at school:

Sample space S = the set of all possible real numbers

An event whose occurrence is inevitable when an experiment is performed is called as Certain event or Sure event.

An event which can never occur when an experiment is performed is called an Impossible event. Events may be 'simple' or 'compound'.

An event is called simple if it corresponds to a single possible outcome of the experiment otherwise it is called compound event or composite event.

For example: In drawing cards from a pack of 52 cards, the chance of getting spade 5 is a simple event and the chance of getting a king is compound event. Occurrence of getting two 8 diamond cards is an impossible event.

Favorable Events: The number of cases favorable to an event in a trial is the number of outcomes which entail the happening of the event.

Equally likely events: The outcomes are said to be equally likely if none of them is expected to occur in preference to other. Thus two or more events are said to be equally likely if each one of them has an equal chance of happening.

For example : when a coin is thrown, the head is as likely to turn up as tail. Hence H and T are equally likely events.

Mutually exclusive events or Incompatible events: If event A happens, then event B cannot, or vice-versa. The two events "it rained on Tuesday" and "it did not rain on Tuesday" are mutually exclusive events. Both cannot happen in a single trial or we can say that the occurrence of any one of them excludes the occurrence of other.

Formally, two events A and B are mutually exclusive if and only if $A \cap B = \varphi$

Exhaustive events: Outcomes are said to be exhaustive when they include all possible outcomes. For example, while rolling a die, the possible outcomes are 1,2,3,4,5 and 6 and hence the exhaustive number of cases is 6.

Independent Events: Two events are independent if the occurrence of one of the events gives us no information about whether or not the other event will occur; that is, the events have no influence on each other.

For example, if a coin is thrown twice, the result of the second throw is no way affected by the result of the first throw. Thus the events are independent events.

NOTES

Dependent events: Two events are said to be dependent if the occurrence or nonoccurrence of an event in any trial affects the occurrence of the other event in other trials.

Complementary events: A is called complementary event of B if A and B are mutually exclusive and exhaustive. When a die is thrown, occurrence of an even number and odd number are complementary events.

1.3.2 Probability of an event :

Classical definition: The outcomes of a random experiment are termed as events.

The probability for the occurrence of an event A is defined as the ratio between the number of favorable outcomes for the occurrences of the event and the total number of possible outcomes.

$$P(A) = \frac{\text{No: of favorable cases}}{\text{Total no: of cases}}$$

Also the probability of non-happening of event A is

$P(\bar{A}) = 1 - m/n$

i.e, $P(\bar{A}) = 1 - P(A)$.

Note: If $P(A) = 1$ the event A is called a certain event and if $P(A) = 0$, the event is called an impossible event.

Also $P(A) + P(\bar{A}) = 1$

Examples

1. The probability of drawing a spade from a pack of 52 well-shuffled playing cards is $13/52 = 1/4 = 0.25$ since

   event E = 'a spade is drawn';

   the number of outcomes corresponding to E = 13 (spades);

   the total number of outcomes = 52 (cards).

1.3.2.1. Statistical definition:

If a trial results in 'n' cases and m of them are favorable to the happening of the event then

$$P(A) = \lim_{n \to \infty} (m/n)$$

***1.3.2.2 Axiomatic approach to probability*:**

Definition of probability :

Let S be a sample space associated with an experiment. To each event A, there is a real number P(A) associated , called the probability of A satisfying the following ***axioms***:

NOTES

i) $P(A) \overset{24}{12} 0$

ii) $P(S) = 1$

iii) If $A_1, A_2, A_3, ------A_n$ are mutually exclusive events, then

$P(A_1 \cup A_2 \cup A_3 ------ \cup An) = P(A_1) + P(A_2) + --------P(A_n)$

Theorem : The probability of an impossible event is zero, i.e, if Φ is the subset (event) containing no sample point , $P(\Phi) = 0$

Proof – The certain event S and the impossible event Φ are mutually exclusive

Hence $P(S \cup \Phi) = P(S) + P(\Phi)$ [Axiom (iii)]

But $S \cup \Phi = S$

Thus $P(S) = P(S) + P(\Phi)$

Thus $P(\Phi) = 0$

Theorem : If $\bar{A}$ is the complimentary event of A, $P(\bar{A}) = 1 - P(A) \overset{c}{\lambda} 1$

Proof : A and $\bar{A}$ are mutually exclusive events, such that $A \cup \bar{A} = S$

Thus $P(A \cup \bar{A}) = P(S)$

$= 1$ [Axiom (ii)]

i.e. $P(A) + P(\bar{A}) = 1$ [Axiom (iii)]

Thus $P(\bar{A}) = 1 - P(A)$

Since $P(A) \overset{24}{12} 0$, It follows that $P(\bar{A}) \overset{c}{\lambda} 1$.

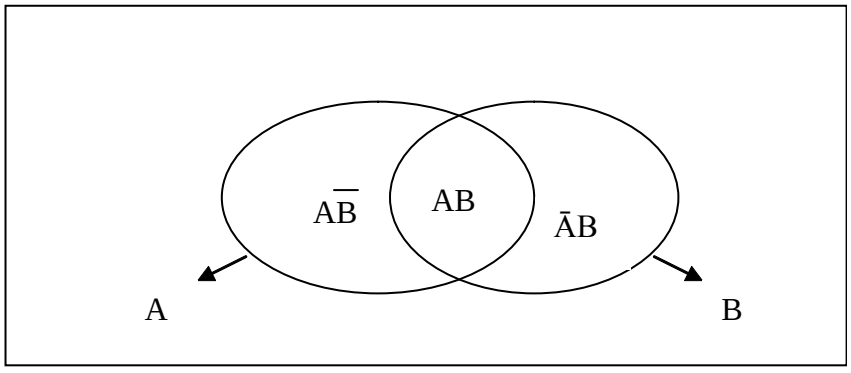
Theorem : If A and B are any 2 events,

$P(A \cup B) = P(A) + P(B) - P(A1B) = P(A) + P(B)$

Proof : A is the union of the mutually exclusive events $A\overline{B}$ and AB and B is the union of the mutually exclusive events $\bar{A}B$ and AB

Thus $P(A) = P(A\overline{B}) + P(AB)$ [Axiom (iii)]

and $P(B) = P(\bar{A}B) + P(AB)$ [Axiom (iii)]



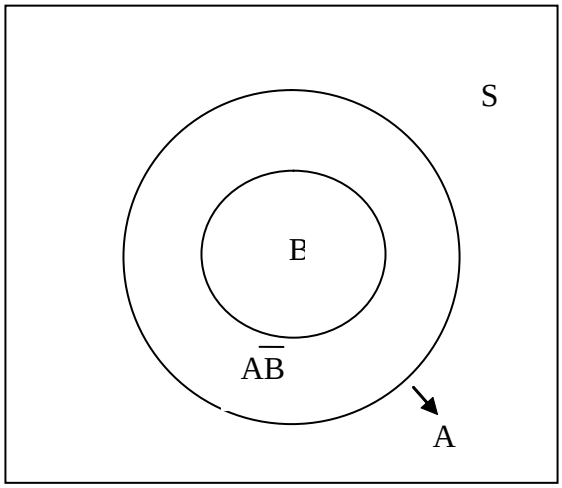
NOTES

Thus P(A) + P(B) = [P(AB) + P(AB) + P($\bar{A}$B) + P(AB)]
= P(A ∪ B) + P (A1B)

This result follows. Clearly, P(A) +P(B) – P (A1B) = P(A) + P(B)

Theorem : If B ⊂ A, P(B) ≤ P(A)

Proof:



B and AB are mutually exclusive events such that B ∪ AB = A

Thus P(B ∪ A B) = P(A)

i.e. P(B) + P AB) = P(A) { Axiom (iii)]

Thus P(B) = P(A)

Note: The above theorem can be extended as

P(A ∪ B ∪ C) = P (at least one A, B anc C occurs)

= P(A) +P(B) + P(C) - P (A1B) - P (B1C) - P (C1A) + P (A1B1C)

1.3.3 Laws of Probability

1.3.3.1 Addition Rule :

The addition rule is a result used to determine the probability that event A or event B occurs or both occur.

The result is often written as follows, using set notation:

P(A ∪ B) = P(A) +P(B) – P(A1B)

where:

P(A) = probability that event A occurs

NOTES

P(B) = probability that event B occurs

P(A ∪ B) = probability that event A or event B occurs

P(A∩B) = probability that event A and event B both occur

For mutually exclusive events, that is events which cannot occur together:

P(A∩B) = 0

The addition rule therefore reduces to

P(A∪B) = P(A) + P(B)

For independent events, that is events which have no influence on each other:

P(A∩B) = P(A)P(B)

The addition rule therefore reduces to

P(A ∪ B) = P(A) +P(B) – P(A)P(B)

Example

Suppose we wish to find the probability of drawing either a king or a spade in a single draw from a pack of 52 playing cards.

We define the events A = 'draw a king' and B = 'draw a spade'

Since there are 4 kings in the pack and 13 spades, but 1 card is both a king and a spade, we have:

P(A ∪ B) = P(A) +P(B) – P(A∩B)

= 4/52 + 13/52 - 1/52 = 16/52

So, the probability of drawing either a king or a spade is 16/52 (= 4/13).

Theorem: Additive law of probability:

If A and B are any two events (subsets of sample space S) are not disjoint, then

P(A ∪ B) = P(A) +P(B) – P(A∩B)

Proof: We have A ∪ B = A ∪ ($\bar{A}$∩B)

Since A and ($\bar{A}$∩B) are disjoint,

P(A ∪ B) = P(A) + P($\bar{A}$∩B)

= P(A) +[P($\bar{A}$∩B) + P(A∩B)] – P(A∩B)

= P(A) +P[($\bar{A}$∩B) ∪ (A∩B)] – P(A∩B)

= P(A) + P(B) – P(A∩B)(since $\bar{A}$∩B and A∩B are disjoint)

NOTES

1.3.3.2 Multiplication Rule

The multiplication rule is a result used to determine the probability that two events, A and B, both occur.

The *Conditional Probability* of an event B, assuming that the event A has happened, is denoted by P(A/B) and defined as

P(B/A) = P(A1B) / P(A)

The multiplication rule follows from the definition of conditional probability.

The result is often written as follows, using set notation:

P(A1B) = P(A/B)P(B) OR P(A1B) = P(B/A)P(A)

where:

P(A) = probability that event A occurs

P(B) = probability that event B occurs

P(A1B) = probability that event A and event B occur

P(A | B) = the conditional probability that event A occurs given that event B has occurred already

P(B | A) = the conditional probability that event B occurs given that event A has occurred already

For *independent* events, that is events which have no influence on one another, the rule simplifies to:

P(A1B) = P(A)P(B) OR P(A and B) = P(A).P(B)

That is, the probability of the joint events A and B is equal to the product of the individual probabilities for the two events.

Theorem: Multiplication Law of Probability or Theorem of compound probabilities.

For two events A and B, P(A1B) = P(A/B)P(B) , P(A)>0

= P(B/A)P(A), P(B)>0

Proof : We have

NOTES

$P(A) = \frac{n(A)}{n(S)}$, $P(B) = \frac{n(B)}{n(S)}$ and $P(A \cap B) = \frac{n(A \cap B)}{n(S)}$ ————————(1)

For the conditional event A/B, favorable outcomes be one of the simple points of B, i.e, for the event A/B, the sample space is B and out of the n(B) of sample points, $n(A \cap B)$ pertain to the occurrence of event A. Hence

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

Rewriting (1) as $P(A \cap B) = \frac{n(B)}{n(S)} \cdot \frac{n(A \cap B)}{n(B)}$

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

Similarly we can prove $P(A \cap B) = \frac{n(A)}{n(S)} \cdot \frac{n(A \cap B)}{n(A)}$

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

Note:

Conditional probabilities P(B/A)and P(A/B) are defined if and only if $P(A) \neq 0$ and $P(B) \neq 0$ respectively.

$P(A/A) = 1$

If A and B are independent, then P(A/B) = P(A) and P(B/A) = P(B).

If A and B are independent their compliments are also independent.

Theorem: If the events A and B are independent, the events $\bar{A}$ and B (and similarly A and $\bar{B}$)are also independent.

Proof:

The events $A \cap B$ and $\bar{A} \cap B$ are mutually exclusive such that $(A \cap B) \cup (\bar{A} \cap B) = B$

$P(A \cap B) + P(\bar{A} \cap B) = P(B)$ by addition theorem

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

$$= P(B) - P(A)P(B) \text{ by product theorem}$$

$$= P(B)[\, 1 - P(A)\,]$$

$$= P(B)P(\bar{A})$$

Based on the above discussed theory some problems are illustrated. Most of the examples are dealing with real life experiments so that it will be interesting for you .

NOTES

Example 1: An integer is chosen from 2 to 15 . What is the probability that it is prime?

Solution: Since there are 14 numbers from 2 to 15 the total number of numbers is 14. Out of 14 numbers 2,3,5,7,11,13 are prime numbers. Hence the number of favourable numbers is 6. Hence the probability that the number chosen is prime is

$$\frac{\text{No: of favorable cases}}{\text{Total no: of cases}} = \frac{6}{14} = \frac{3}{7}$$

Example 2: What is the chance that a) leap year selected at random will contain 53 Sundays b)a non-leap year selected at random will contain 53 Sundays?

Solution:

a) In a leap year, there are 366 days i.e, 52 weeks and 2 days, remaining 2 days can be

any two days of the week. The different possibilities are: Sunday and Monday, Monday and Tuesday, Tuesday and Wednesday, Wednesday and Thursday, Thursday and Friday, Friday and Saturday, Saturday and Sunday.

In order to have 53 Sundays, out of remaining two days one must be Sunday.

No: of favorable cases = 2

Total no: of cases = 7

Required Probability = 2/7

b) In a non leap year, there are 365 days i.e, 52 weeks and 1 day. The remaining 1 day

can be any day of the week.

Total no: of cases = 7

There will be 53 Sundays if the remaining one day is Sunday.

No: of favorable cases = 1

Required probability = 1/7.

Example 3: If we draw a card from a well shuffled pack of 52 cards, what is the probability that the card is either an ace or a king?

Solution:

The events that the card drawn is either an ace or a king are mutually exclusive. Let the event of the ace drawn be denoted by A and the king drawn be denoted by K. A pack has 4 aces and a kings.

Thus P(A) = 4/52 and P(K) = 4/52

The probability that the card is either ace or a king =

NOTES

P(A or K) = P(A ∪ K) = P(A) + P(K) = 4/52 + 4/52 = 2/13.

Example 4: A,B and C toss a coin. The one who get the head first wins. What are their respective chances?

Solution: The probability that A may win in the first round = ½. He may win in the second round, after all of them failed in the first round, with the probability = (½.½ ½). ½ = $(½)^3$½, iin the third round with probability = $(½)^6$ ½..and so on.

Now by addition theorem the chance of A's success = ½ + $(½)^3$½ + $(½)^6$ ½ + - - - - which is a G P with first term ½ and common ratio = $(½)^2$.

Therefore the above expression is = $\frac{. ½ .}{1 - (½)^3}$ = 4/7.

B may win the first round after A's failure with probability = (½)(½)= ¼, the second round with probability = $(½)^3$ ¼ and so on.

Therefore B's chances of success = ¼ +¼ $(½)^3$ + - - - = $\frac{. ¼ .}{1 - (½)^3}$ = 2/7

Similarly, the chance of winning the game for C = $\frac{. 1/8 .}{1 - (½)^3}$ = 1/7

Example 5: If from a lottery of 30 tickets numbered 1, 2, 3, - - - - 30 four tickets are drawn, what is the chance that those marked 1 and 2 are among them?

Solution:

Out of 30 tickets 4 tickets can be chosen in $30C_4$ ways. In the 4 tickets 1 and 2 should be present. The remaining 2 tickets should be chosen from tickets numbered 3, 4, - - - -30 i.e. 28tickets which can be done in $28C_2$ ways. Thus the required probability is

$\frac{28C_2}{30C_4} = \frac{2}{145}$

Example 6: A problem in mathematics is given to five students A_1, A_2, A_3, A_4 and A_5. Their chances of solving it are 1/6, 1/5, ¼, 1/3, ½ respectively. What is the probability that the problem will be solved?

Solution:

The probability that A_1 fails to solve the problem = 1 – 1/6 = 5/6.

The probability that A_2 fails to solve the problem = 1 – 1/5 = 4/5.

The probability that A_3 fails to solve the problem = 1 – 1/4 = 3/4.

The probability that A_4 fails to solve the problem = 1 – 1/3 = 2/3.

The probability that A_5 fails to solve the problem = 1 – 1/2 = 1/2.

NOTES

The probability that the problem is not solved by all five students

$$= \frac{5}{6}\ \frac{4}{5}\ \frac{3}{4}\ \frac{2}{3}\ \frac{1}{2} = \frac{1}{6}$$

Therefore the probability that the problem will be solved $1 - \frac{1}{6} = \frac{5}{6}$

Example 7: A lot consists of 10 good articles, 4 with minor defects and 2 with majoe defects. Two articles are chosen from the lot at random (without replacement). Find the probability that i) both are good, ii)both have major defects, iii) at least 1 is good, iv) at most 1 is good v)exactly 1 is good vi)neither has major defects and vii) neither is good?

Solution:

Since the articles are drawn without replacement, we can consider that both the articles are drawn simultaneously.

i) P(both are good) = $\frac{\text{No : of ways of drawing 2 articles from good articles}}{\text{Total no : of ways of drawing 2 articles.}}$

$$= \frac{10C_2}{16C_2} = \frac{3}{8}$$

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

$$= \frac{2C_2}{16C_2} = \frac{1}{120}$$

iii) P(atleast 1 is good) = P (exactly 1 is good and 1 is bad or both are good)

$$= \frac{10C_1 \times 6C_1 + 10C_2}{16C_2} = \frac{7}{8}$$

iv)P(atmost 1 is good) = P(none is good or 1 is good and 1 is bad)

$$= \frac{10C_0 \times 6C_2 + 10C_1 \times 6C_1}{16C_2} = \frac{5}{8}$$

v) P(exactly 1 is good) = P(1 is good and 1 is bad)

$$= \frac{10C_1 \times 6C_1}{16C_2} = ½$$

NOTES

vi) P(neither has major defects) = P(both are non-major defective articles)

$$= \frac{14C_2}{16C_2} = \frac{91}{120}$$

vii)P(neither is good) = P(both are defective)

$$= \frac{6C_2}{16C_2} = \frac{1}{8}$$

Example 8: A and B throw alternately with a pair of dice. A wins if he throws (sum of numbers on top two faces) 6 before B throws 7, and B wins if he throws 7 before A throws 6. If A begins, find his chance of winning.

Solution:

The sum 6 can he got in 5 ways.[(1,5), (2,4),(3,3)(4,2)(5,1)]

The probability of A throwing 6 is 5/36.

Therefore the probability A throwing 6 = 1 – 5/36 = 31/36

Similarly, the probability of B throwing 7 is 6/36.

Therefore the probability of B not throwing 7 = 1 – 6/36 = 5/6.

Now A can win if he throws 6 in the 1st, 3rd , 5th , - - - - -throws.

Hence the chance of A winning

$$= \frac{5}{36} + \frac{31}{36}\frac{5}{6}\frac{5}{36} + \frac{31}{36}\frac{5}{6}\frac{31}{36}\frac{5}{6}\frac{5}{36} + \text{- - - -}$$

$$= \frac{5}{36}\left[1 + \frac{31.5}{36.6} + \left(\frac{31.5}{36.6}\right)^2 + \left(\frac{31.5}{36.6}\right)^3 + \text{- - - - -}\right]$$

$$= \frac{5}{36}\left(\frac{1}{1 - \frac{31.5}{36.36}}\right) = \frac{30}{61}$$

Example 9: A box contains 4 bad and 6 good tubes. Two are drawn out from the box at a time. One of them is tested and found to be good. What is the probability that the other one is also good?

Solution:

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

NOTES

P(A1B) = P(both tubes drawn are good)

$$= \frac{6C_2}{10C_2} = \frac{1}{3}$$

With the condition that one is good, we have to find the conditional probability that the other tube is also good is required. i.e P(B/A) is required.

$$P(B/A) = \frac{P(A1B)}{P(A)} = \frac{1/3}{6/10} = \frac{5}{9}$$

Example 10: The odds against the student X solving a problem in mathematics are 8:6 and odds in favour the student Y solving the same problem are 14:16 i) What is the chance that the problem will be solved if they both try? ii) what is the probability that both working independently will solve the problem? Iii) What is the probability that neither solves the problem?

Solution:

Let A be the event that the student X solves the problem, and B be the evnt that the student Y solves the problem. Then by data

$P(\bar{A}) = 8/14$ and $P(B) = 14/30$.

Thus, $P(A) = 1 - P(\bar{A}) = 1 - 8/14 = 6/14$

$P(\bar{B}) = 1 - P(B) = 1 - 14/30 = 16/30$

i)Probability that the problem will be solved

= P(any one solves the problem)

= P(A or B)

= P(A) +P(B) – P(A1B) = P(A) +P(B) – P(A)P(B)

= 6/14 + 14/30 – (6/14 x 14/30)

ii) Probability of solving the problem if they work independently is P(A and B)

= P(A1B) = P(A)P(B) = 6/14 x14/30

iii) Probability that both will not solve the problem is

= P(A and B) = P(A)P(B) = 1/14 x 16/30

1.3.4 Theorem of total probability

If $B_1, B_2, \ldots B_n$ be a set of exhaustive and mutually exclusive events, and A is another associated with B_i, then

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

NOTES

1.3.5 Baye's Theorem or Theorem of probability of causes

If B_1, B_2, - - - Bn be a set of exhaustive and mutually exclusive events associated with a random experiment and A is another event associated with B_i, then

$$P(B_i / A) = \frac{P(B_i)P(A/B_i)}{\left(\sum_{i=1}^{n} P(B_i)P(A/B_i)\right)}, \quad i = 1,2,3, - - n$$

Proof:

$P(B_i \cap A) = P(B_i) \times P(A/B_i) = P(A) \times P(B_i / A)$

$$P(B_i / A) = \frac{P(B_i) \times P(A/B_i)}{P(A)}$$

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

Example 1: Bolts are manufactured by three machines A, B, C. A turns out twice as many bolts as B and machines B and C produce equal number of bolts. 2% of bolts produced by A and by B are defective and 4% of bolts produced by C are defective. All bolts are put into one stock pile and one is chosen from this pile. What is the probability that it is defective?

Solution:

Let E_1, E_2, E_3 be the events that the bolts are manufactured by machines A, B,C respectively.

Hint: given A turns out twice as many bolts as B and machines B and C produce equal number of bolts.

$P(A) = 2P(B)$ & $P(B) = P(C)$

But $P(A) + P(B) + P(C) = 1$

- ⇨ $2P(B) + P(B) + P(B) = 1$
- ⇨ $4P(B) = 1$
- ⇨ $P(B) = ¼$
- ⇨ $P(C) = ¼$
- ⇨ $P(A) = ½$

NOTES

Therefore $P(E_1) = ½$, $P(E_2) = ¼$, $P(E_3) = ¼$

Let D be the event of chosing a defective bolt.

$P(D/E_1)$ = Probability of choosing a defective bolt from machine A = 2/100

$P(D/E_2)$ = Probability of choosing a defective bolt from machine B = 2/100

$P(D/E_3)$ = Probability of choosing a defective bolt from machine C = 4/100

Probability of choosing a defective bolt

$$P(A) = \sum_{i=1}^{3} P(E_i)\, P(A/E_i) \text{ by total probability theorem}$$

$$= \frac{1}{2}\frac{2}{100} + \frac{1}{4}\frac{2}{100} + \frac{1}{4}\frac{4}{100} = \frac{1}{40}$$

Example 2: The contents of urns I, II, and III are as follows: 2 white, 3blacks and 4 red balls; 3 white , 2 black and 2 red balls and 4 white, 1 black and 3 red balls. An 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 urns I, II, and III?

Solution:

Let E_1, E_2 and E_3 denote the events that the urn I, II, and III be chosen respectively and let A be the event that the two balls taken from the urn be white and red.

$P(E_1) = P(E_2) = P(E_3) = 1/3$

$$P(A/E_1) = \frac{2C_1 \times 4C_1}{9C_2} = 2/9$$

$$P(A/E_2) = \frac{3C_1 \times 2C_1}{7C_2} = 2/7$$

$$P(A/E_3) = \frac{4C_1 \times 3C_1}{8C_2} = 3/7$$

Now we have to calculate $P(E_1/A)$, $P(E_2/A)$ and $P(E_3/A)$

$$P(E_1 / A) = \frac{P(E_1)P(A / E_1)}{\sum_{i=1}^{3} P(E_i)P(A/E_i)},$$

NOTES

$$= \frac{(1/3)(2/9)}{(1/3.2/9) + (1/3.2/7) + (1/3.3/7)}$$

$$P(E_2 / A) = \frac{P(E_2)P(A / E_2)}{\sum_{i=1}^{3} P(E_i)P(A / E_i)},$$

$$= \frac{(1/3)(2/7)}{(1/3.2/9) + (1/3.2/7) + (1/3.3/7)}$$

$$P(E_3 / A) = 1 - (P(E_1 / A) + P(E_2 / A))$$

Example 3: An urn contains5 white and 3 green balls and another urn contains 3 white and 7 green balls. Two balls are chosen at random from the first urn and put into the second urn . Then a ball is drawn from the second urn. What is the probability that it is a white ball?

Solution:

The two balls drawn from the first urn can be

i)both white which is denoted byE_1

ii) both green which is denoted byE_2

iii)one white and one green which is denoted by E_3

$$P(E_1) = \frac{5C_2}{8C_2} = \frac{5}{14}$$

$$P(E_2) = \frac{3C_2}{8C_2} = \frac{3}{28}$$

$$P(E_3) = \frac{5C_1 3C_1}{8C_2} = \frac{15}{28}$$

After the balls are transferred from the first urn to the second urn, the second urn will contain

i) 5 white and 7 green balls

ii) 3 white and 9 green balls

iii) 4 white and 8 green balls

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

NOTES

$$P(A / E_1) = \frac{5C_1}{12C_1} = \frac{5}{12}$$

$$P(A / E_2) = \frac{3C_2}{12C_2} = \frac{3}{12}$$

$$P(A / E_3) = \frac{4C_1}{12C_2} = \frac{4}{12}$$

Required probability of choosing a white ball = P (A)

$$= \sum_{i=1}^{3} P(E_i)\, P(A / E_i) \text{ (by total probability theorm)}$$

$$= \frac{5}{14} \times \frac{5}{12} + \frac{3}{28} \times \frac{3}{12} + \frac{15}{28} \times \frac{14}{12}$$

$$= \frac{125}{336} = 0.372$$

Example 4: A toy is rejected if the design is faulty or not. The probability that the design is faulty is 0.1 and that the toy is rejected if the design is faulty 0.95 and otherwise 0.45. If a toy is rejected, what is the probability that it is due to faulty design?

Solution:

Let D_1, D_2 denote the events that the design is faulty or not. Let A denote the event that the toy is rejected.

$P(D_1) = 0.1 \quad P(D_2) = 1 - 0.1 = 0.9$

$P(A/D_1) = 0.95$ and $P(A/D_2) = 0.45$

P(of rejection due to faulty design) =

$$P(D_1/A) = \frac{P(D_1)\, P(A/D_1)}{P(D_1)\, P(A/D_1\, P(D_2)\, P(A/D_2)}$$

$$= \frac{0.1 \times 0.95}{0.1 \times 0.95 + 0.9 \times 0.45} = 0.19$$

Example 5: For a certain binary communication channel, the probability that a transmitted '0' is received as a '0' is 0. 95 and the probability that a transmitted '1' is received as '1' is 0.90. If the probability that a '0' is transmitted is 0.4, find the probability that

NOTES

i) a '1' is received and

ii) a '1' was transmitted given that a '1' was received

Solution :

Let A = the event of transmitting '1'. $\bar{A}$ = the event of transmitting'0',

B = the event of receiving '1' and B = the event of receiving '0'.

Given: $P(\bar{A}) = 0.4$, $P(B/A) = 0.9$ and $P(B/\bar{A}) = 0.95$

Thus $P(A) = 0.6$ and $P(B/\bar{A}) = 0.05$

By the theorem of total probability

$P(B) = P(A) \times P(B/A) + P(\bar{A}) \times P(B/\bar{A})$

$= 0.6 \times 0.9 + 0.4 \times 0.05$

$= 0.56$

By Baye's theorem $P(A/B) = \frac{P(A) \times P(B/A)}{P(B)} = \frac{0.6 \times 0.9}{0.56} = \frac{27}{28}$

1.3.6 Bernoulli's Trials

Let us consider n independent repetitions of a random experiment. If A is an event associated with E such that P(A) remains the same for the repetitions, the trials are called Bernoulli's trial.

Theorem: If the probability of occurrence of an event in a single trial of Bernoulli's experiment is p, then the probability that the event occurs exactly x times out of n independent trials is equal to $nCr\, q^{n-x}p^{x}$, where $q = 1 - p$, the probability of failure of the event.

Example: A die is tossed until 6 appears. What is the probability that it must be tossed more than 4 times.

Solution:

$P(X = x) = (1/6)\,(5/6)^{x-1}$ $x = 1,2,3, - - - -$

$P(x>4) = 1 - P(x = 4)$

$= 1 - \sum_{x=1}^{4} (1/6)\,(5/6)^{x-1}$

$= 1 - (1/6)[1 + 5/6 + 25/36 + 125/216] = 0.48225$

Have you understood ?

1) Which of the following is an experiment?

NOTES

a) Tossing a coin.
b) Rolling a single 6-sided die.
c) Choosing a marble from a jar.
d) All of the above

2) Which of the following is an outcome?

a) Rolling a pair of dice.
b) Landing on red.
c) Choosing 2 marbles from a jar.
d) None of the above.

3) Which of the following experiments does NOT have equally likely outcomes?

a) Choose a number at random from 1 to 7.
b) Toss a coin.
c) Choose a letter at random from the word SCHOOL.
d) None of the above

4) A number from 1 to 11 is chosen at random. What is the probability of choosing an odd number?

a) $\frac{1}{11}$

b) $\frac{5}{11}$

c) $\frac{6}{11}$

d) None of the above.

Answers: 1) d. 2)b. 3)c. d)c.

Short answer questions:

1. What is random experiment? Give an example.
2. Give the axiomatic definition of probability.
3. State axioms of probability.
4. State addition theorem as applied to any 2 events .Extend it to any three events.
5. In a random experiment P(A) =1/12,P(B) = 5/12 and P(B/A) = 1/15 find P (AUB) (Solution: 89/180)
6. If P(A) = 0.5, P(B) = 0.3 and P(A1B) = 0.15, find P(A/ B)
7. State the theorem of total probability.
8. Satate Baye's theorem.

TRY YOURSELF !

NOTES

1. From a bag containing 3 red and 2 black balls, 2 balls are drawn at random. Find the probability that they are of the same colour?

(Solution:2/5)

2. Event A and B are such that P(A + B) =3/4, P(AB) = ¼ and P($\bar{A}$) = 2/3 find P(B) (Solution:2/3)

3. In a random experiment P(A) = 1/12, P(B) = 5/12 and P(B/A) = 1/15. Find P (AUB) (Solution: 89/180)

4. If P(A) = 0.5, P(B) = 0.3 and P(A1B) = 0.15, find P(A/ B)

(Solution: 0.5)

5. Probability that India wins a cricket match against Australia is known to be 2/5 If India and Australia play 3 test 3 matches what is the probability that i) India will loose all the three matches ii) India will win all the tests iii) India will win atmost one match.

(Solution: i) 27/125 ii) 98/125 iii) 8/125 iv) 81/125)

6. Two weak students attempt to write a program. Their chances of writing the program successfully is 1/8and 1/12 and the chance of making a common error is 1/10001.Find the chance that the program is correctly written.

(Solution: 0.9924)

7. A fair dice is rolled 5 times. Find the probability that 1 shows twice, 3 shows twice and 6 shows once. (Solution: 0.0039)

8. One integer is chosen at random from the numbers 1,2,3,....100.what is the probability that the chosen number is divisible by (i)6 or 8and (ii) 6 or 8or both.

(Solution: 1/5 , 6/25)

9. Urn I has 2 white and 3 black balls, urn II contains 4 white and 1 black balls and urn III contains 3 white and 4 black balls. An urn is selected at random and is found to be white. Find the probability that urn I was selected.

(Solution: 14/57)

1.4 RANDOM VARIABLES

You have seen a number of examples of sample spaces associated with various experiments. In some, the outcomes were numerical and in some others the outcomes were non numerical.

NOTES

For example in the experiment concerned with tossing a dice, we may have any one of the following outcome as 1,2,3,4,5 and 6 which is numerical in nature. While the result of a coin tossing experiment in which a coin is tossed once, we have the outcome and head or tail which is non numerical in nature.

As it is often convenient to describe the outcome of a random experiment by a number, we will assign a number to each non numerical outcome of the experiment..
For example in the coin tossing experiment, we can assign the value of 0 to the outcome of getting heads and 1 to the outcome of getting tails.
Thus in any experimental situation we can assign a real number x to every element s of the sample spaces.

Random Variable – Let E be an experiment and S a sample space associated with the experiment. A function X assigning to each element s ε S, a real number X is called a random variable.
The set of values which the random variable X takes is called the spectrum of the random variable.

1.4.1 Types of Random Variable –

a) Discrete Random Variable
b) Continuous Random Variable

1.4.1.1 Discrete Random Variable – A random variable is said to be discrete if it assumes only a finite or countably infinite values of X, that is the range space R contains a finite or countably infinite points. The possible values of x may be listed as x_1, x_2 ….

In the finite cases the list terminates. In the countably infinite cases, the list continues.

Let x_1, x_2 …. be possible values of a discrete random variable X. Then $P(x_i)$ is called the probability function or probability mass function or point probability function of the discrete random variable X if

i) $P(x_i) \overset{24}{12} 0$ or i=1,2……
ii) $\sum_i P(x_i) = 1$

The collection of pairs (x_i, $P(x_i)$), i=1,2…. is called the probability distribution.

1.4.1.2 Continuous Random Variable – A random variable X is called a continuous random variable if X takes all its possible values of an interval (or) equivalently. If the range space R_x of the random variable X is an interval or a collection of intervals, X is called continuous random variable.

NOTES

Probability Density function (p.d.f)

If x is a continuous random variable such that P{ x – dx/2 = X = x + dx/2 } = f(x)dx, then f(x) is called the probability density function (p.d. f) of X provided f(x) satisfies the following conditions

i) $f(x) \geq 0$ for all $x \in R_x$

ii) $\int_{R_x} f(x)\,dx = 1$

Where R_x is the range of X

1.4.1.3 Properties

1) $f(x) >= 0$ for all $x \in R_x$

2) $\int_{-\infty}^{\infty} f(x)\,dx = 1$ or $\int_{R_x} f(x)\,dx = 1$

3) $P(a < x < b) = \int_a^b f(x)\,dx$

4) Probability at a particular point is zero. i.e it is impossible that a continuous random variable assumes a specific value since, $P(X = a) = P(a \leq X \leq a) = \int_a^a f(x)\,dx = 0$

This means that it is almost impossible that a continuous random variable assumes a specific value. Hence

$P(a \leq X \leq b) = P(a \leq X < b) = P(a < X \leq b) = P(a < X < b)$

1.4.2 Cumulative Distribution function (c.d.f)

If X is a random variable, discrete or continuous, then $P(X \leq x)$ is called the cumulative distribution function of X or distribution fn of X and denoted a F(x)

If X is discrete, $F(x) = \sum_{x_j \leq x} P_j$

If X is Continous, $F(x) = \int_{-\infty}^{x} f(x)\,dx$

NOTES

1.4.2.1 Properties

1) F(x) is a non decreasing function of x, i.e if $x_1 < x_2$ then $F(x_1) = F(x_2)$

2) $F(-\infty) = 0$ and $F(\infty) = 1$

3) $P(a \leq X \leq b)$ can also be expressed in terms of distribution function as
$P(a \leq X \leq b) = F(b) - F(a)$

4) If F(x) is a Cumulative Distribution function of a continuous random variable X then the Probability Density function of X, f (x) is given by

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

5) If X is a discrete random variA1Ble taking values $x_1, x_2 \ldots$
.where $x_1 < x_2 < x_3 \ldots < x_{i-1} < x_i < \text{- - -}$ then $P(X = x_i) = F(x_i) - F(x_{i-1})$

Example1: The probability function of a random variable X is given by
P(x) = ¼ for x = 2
= ¼ for x = 0
= ½ for x = 10.
Verify that the total probability is 1.Evaluate the following probabilities
a)$P(X \leq 0)$ b)$P(X<0)$ c) $P(|X| \leq 2)$ and d) $P(0 \leq X \leq 10)$.

Solution:

$$\sum_{j=1}^{\infty} p_j = ¼ + ¼ + ½ = 1$$

Hence the total probability is 1

a) $P(X \leq 0) = P(X= -2) + P(X= 0) = ¼ + ¼ = ½$

b) $P(X<0) = P(X= -2) = ¼$.

c) $P(|X| \leq 2) = P(-2 \leq X \leq 2) = P(X= -2) + P(X= 0) = ½$.

d) $P(0 \leq X \leq 10) = P(X= 0) + P(X= 10) = ¾$.

Example 2: Consider a random experiment of tossing three times a fair coin. Let X denote the number of tails and Y denotes the number of consecutive tails.

Find
i)The probability distribution of X and Y
ii)the distribution function of x
iii)the probability distribution of X + Y and XY.

NOTES

Solution:
The sample space of the random experiment is

S= {HHH, HHT, HTH,THT,HTT,THH,TTH,TTT}

Each element of S occurs with probability 1/8. The values of X, Y, X + Y and XY for each outcome is tabulated.

Event	HHH	HHT	HTH	HTT	THH	THT	TTH	TTT
X	0	1	1	2	1	2	2	3
Y	0	0	0	2	0	0	2	3
X+Y	0	1	1	4	1	2	4	6
XY	0	0	0	4	0	0	4	9

i)Probability distribution of X

Value of X, x	0	1	2	3
p(x)	1/8	3/8	3/8	1/8

ii)Probability distribution of Y

Value of Y, y	0	2	3
P(y)	5/8	2/8	1/8

iii)distribution function of X

X	[-∞,0)	[0,1)	[1,2)	[2,3)	[3,∞)
F(X)	0	1/8	4/8	7/8	1

NOTES

iv)Probability distribution of X + Y

Value of X+Y	0	1	2	4	6
p(x)	1/8	3/8	1/8	2/8	1/8

v) Probability distribution of XY

Value of XY	2	4	9
p(x)	5/8	2/8	1/8

Example 3: Find the constant c so that the function

$f(x) = cx^2$, $0<x<3$
$= 0$, otherwise is a pdf . Find the distribution function and evaluate $P(1<x<2)$.

Solution:
Given $f(x) = cx^2$, $0<x<3$
$= 0$, otherwise
By the property of pdf

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

$$\int_0^3 cx^2 dx = 1 \Rightarrow [cx^3/3]_0^3 = 1 \Rightarrow 9c = 1 \Rightarrow c = 1/9.$$

The distribution function is given by

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

When $x \le 0$, $F(x) = P(X \le 0) = \int_{-\infty}^{x} f(x)dx = 0$ since f(x) is not defined for $x<0$.

When $0<x<3$, $F(x) = P(X \le x) = \int_{-\infty}^{0} f(x)dx + \int_{0}^{x} f(x)dx = \int_{0}^{x} x^2/9\,dx = x^3/27]_0^x = x^3/27.$

When x=3

$$F(x) = P(X \le x) = \int_{-\infty}^{0} f(x)dx + \int_{0}^{3} f(x)dx + \int_{3}^{x} f(x)dx = \int_{0}^{3} x^2/9\,dx = x^3/27]_0^3 = 1$$

NOTES

$$F(x) = \begin{cases} 0, & x=0 \\ x^3/27, & 0<x<3 \\ 1, & x=3. \end{cases}$$

$$P(1<x<2) = \int_1^2 f(x)dx = \int_1^2 x^2/9dx = 1/9(x^3/3)_1^2 = 7/27.$$

Example 4: A random variable has the following probability distribution

X	:	0	1	2
p(x)	:	$3c^2$	$4c - 10c^2$	$5c - 1$

Find i)the value of c ii) P(0<X<2/X>0) and iii) the distribution function of X iv)the largest value of X for which F(x) < ½ and v) smallest value of X for which F(x) > ½

Solution:

i)since $\Sigma p(x) = 1$, $3c^2 + 4c - 10c^2 + 5c - 1 = 1$

$7c^2 - 9c + 2 = 0$

$c = 2/7, 1$

The value c = 1makes some p(x) negative which is meaningless.

Therefore c = 2/7

ii) P(0<x<2/x>0)

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

$$P(0<x<2/x>0) = \frac{P(0<x<2 \cap x>0)}{P(x>0)} = \frac{4c - 10c^2}{P(x=1) + P(x=2)} = \frac{4c - 10c^2}{4c - 10c^2 + 5c - 1}$$ where c =2/7

$$= \frac{4c - 10c^2}{-10c^2 + 9c - 1} = \frac{8/7 - 40/49}{-40/9 + 18/7 - 1} = 16/37$$

iii)cdf is defined as F(x) = P(X=x)

when x<0 F(x) = 0

when $0 \le x < 1$ F(x) = P(X = 0) = 12/49.

when $1 \le x < 2$ F(x) = P(X = 0) + P(X = 1) = 4/7

when $x \ge 2$ F(x) = P(X = 0) + P(X = 1) + P(X = 2) = 1

iv) form the above cdf it is clear that the largest value of X for which F(x)<½ isx = 0

v) Similarly form the above cdf it is clear that the smallest value of X for which F(x) > ½ is x = 1

NOTES

Example 5: The probability function of an infinite discrete distribution is given by $P(X=j) = 1/2^j (j = 1,2,\ldots 8)$. Verify that the total probability is 1 and find the mean and variance of the distribution. Find also P(X is even). P(X=5) and P(X is divisible by 3).

Solution:

Let $P(X=j) = p_j$

$$\sum_{j=1}^{\infty} p_j = \tfrac{1}{2} + 1/2^2\ 1/2^3 + \text{- - - - - } \infty \text{ which is geometric series}$$

$$= \frac{\tfrac{1}{2}}{1-\tfrac{1}{2}} = 1$$

The mean of X is defined as $E(X) = \sum_{j=1}^{\infty} j\, p_j$

Therefore $E(X) = a + 2a^2 + 3a^3 + \text{- - - - } \infty$ where $a = 1/2$

$= a(1 + 2a + 3a^2 + \text{- - - - } \infty)$

$= a(1 - a)^{-2}$

$$= \frac{\tfrac{1}{2}}{(\tfrac{1}{2})^2} = 2.$$

The variance of X is defined as $V(X) = E(X^2) - [E(X)]^2$

Where $[E(X)]^2 = \sum_{j=1}^{\infty} j^2 p_j$

$[E(X)]^2 = \sum_{j=1}^{\infty} j^2 a^j$, where $a = 1/2$

$$= \sum_{j=1}^{\infty} [j(j+1) - j]\, a^j = \sum_{j=1}^{\infty} j(j+1)\, a^j - \sum_{j=1}^{\infty} j a^j$$

$= a(1.2+2.3a+3.4a^2+ \text{- - - -} \infty) - a(1 + 2a + 3a^3 + \text{- - - } \infty)$

$= a \times 2(1-a)^3 - a \times (1-a)^{-2}$

$$= \frac{2a}{(1-a)^3} - \frac{a}{(1-a)^2} \quad \text{(where } a = 1/2\text{)}$$

$= 8 - 2 = 6$

Variance $= E(X^2) - [E(X)]^2 = 6 - 4 = 2$

NOTES

P(X is even) = P(X = 2 or X = 4 or X= 6 or- - - - - -)

= P(X = 2)+P(X = 4) + P(X= 6) +- - - - - - ∞

$=(½)^2 +(½)^4 +(½)^6 + +- - - - - - \infty$

$= \frac{¼}{1 - ¼}$

= 1/3

P(X=5) = P(X = 5 or X = 6 or X= 7 or- - - - - -)

= P(X = 5)+P(X = 6) + P(X= 7) +- - - - - - ∞

$= \frac{(½)^5}{1- ½}$

= 1/16

P(X is divisible by 3) = P(X = 3 or X = 6 or X= 9 or- - - - - -)

= P(X = 3)+P(X = 6) + P(X= 9) +- - - - - - ∞

$= (½)^3 +(½)^6 +(½)^9 + +- - - - - - \infty$

$= \frac{1/8}{1 – 1/8}$

= 1/7

Example 6: The diameter of an electric cable X is a continuous RV with pdf f(x) = kx(1 -x), 0=X=1. Find i)the value of k, ii)cdf of X, iii) the value of a such that P(X<a) = 2P(X>a) and iv) P(X ≤ ½/1/3<X<2/3)

Solution

i) If f(x) is pdf $\int_0^1 f(x)dx = 1$

$\int_0^1 kx(1 -x)dx = 1 \Rightarrow \int_0^1 k(x –x^2)dx = 1 \Rightarrow k[x^2/2 - x^3/3]_0^1 = 1$

$\Rightarrow k/6 =1 \Rightarrow k =6$

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

When $x \le 0$, $F(x) = P(X \le 0) = \int_{-\infty}^{x} f(x)dx = 0$ since f(x) is not defined for x<0.

NOTES

When $0 \le x \le 1$, $F(x) = P(X \le x) = \int_{-\infty}^{0} f(x)dx + \int_{0}^{x} f(x)dx = \int_{0}^{x} k(x - x^2)\,dx = k[x^2/2 - x^3/3]_0^x =$.

$= 3x^2 - 2x^3$

When $x \le 1$, $F(x) = P(X \le x) = \int_{-\infty}^{0} f(x)dx + \int_{0}^{1} f(x)dx + \int_{1}^{x} f(x)dx = \int_{0}^{1} k(x - x^2)\,dx = k[x^2/2 - x^3/3]_0^1$

$= 1$

Therefore cdf is given by

$$F(x) = \begin{cases} 0, & x \le 0 \\ 3x^2 - 2x^3, & 0 \le x \le 1 \\ 1, & x \ge 1. \end{cases}$$

iii) $P(X<a) = 2P(X>a)$

$$\Rightarrow \int_0^a f(x)dx = 2\int_a^{\infty} f(x)dx$$

$$\Rightarrow \int_0^a f(x)dx = 2\int_a^{1} f(x)dx$$

$$\Rightarrow \int_0^a k(x - x^2)dx = 2\int_a^{1} k(x - x^2)dx$$

$$\Rightarrow k[x^2/2 - x^3/3]_0^a = 2\,k[x^2/2 - x^3/3]_a^1$$

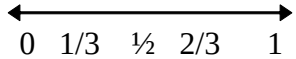
$$\Rightarrow a^2/2 - a^3/3 = 2[1/2 - 1/3 - a^2/2 + a^3/3]$$

$$\Rightarrow 3a^2/2 - 3a^3/3 - 1/3 = 0$$

$$\Rightarrow \frac{9a^2 - 6a^3 - 2}{6} = 0$$

The value of a is the solution of the cubic equation $6a^3 - 9a^2 + 2 = 0$, which lies between o and 1.

iv) $\dfrac{P(0 \le X \le 1/2 \text{ n } 1/3 < X < 2/3)}{P(1/3 < X < 2/3)} = \dfrac{P(1/3 < X < 1/2)}{P(1/3 < X < 2/3)}$



NOTES

$$\Rightarrow \frac{\int_{1/3}^{1/2} k(x-x^2)\,dx}{\int_{1/3}^{1/2} k(x-x^2)\,dx} = \frac{[x^2/2 - x^3/3]_{1/3}^{1/2}}{[x^2/2 - x^3/3]_{1/3}^{2/3}} = \frac{1/8 - 1/24 - 1/18 + 1/81}{4/18 - 8/81 - 1/18 + 1/81}$$

$$= (13/24) / (13/162) = 1/2$$

Example 7: Suppose that the life of a certain radio tube (in hours) is a continuous R V with pdf

$f(x) = 100/x^2$, $x \geq 100$
$= 0$, elsewhere

i) Find the distribution function
ii) If 3 such tubes are inserted in a set what is the probability that none of the three tubs will be replaced during the first 150 hours of operation.
iii) What is the probability that exactly one tube will have to be replaced after 150 hours of service?
iv) What is the probability that all the tubes will be replaced after 150 hours of service?
v) What is the maximum number of tubes that may be inserted into a set so that there is probability of 0.1 that after 150 hour of service all of them are still functioning?

Solution:
The distribution function is given by $P(X \leq x)$

When $x \leq 100$
$F(x) = P(X \leq x) = 0$ since $f(x)$ is 0

When $x > 100$

$$F(x) = P(X \leq x) = \int_{100}^{x} f(x)dx = \int_{100}^{x} 100/x^2\,dx = 100\,[-1/x]_{100}^{x} = 1 - 100/x$$

The distribution function is

$F(x) = 0$, $x \leq 100$
$= 1 - 100/x$, $x > 100$

ii) Let the random variable X be the life of a radio tube.
Probability of a tube will last for 150 hours = $P(X \leq 150)$

$$= \int_{100}^{150} (100/x^2)\,dx = 1/3.$$

NOTES

Therefore the probability that one tube will not be replaced in the first 150 hours is 1/3. The probability of two will not be replaced in the first 150 hours is $(1/3)^2$. Probability that none of the three tubes will have to replaced during first 150 hours is $(1/3)^3$

iii)The probability that a tube will not last for 150 hours = 1-1/3 = 2/3.
Probability that two tubes last for 150 hours = $(1/3)^2$.Since there are $3C_1$ ways of choosing 1 tube from the 3 tubes, the required probability that exactly after 150 hours $= 3C_1(1/3)^2(2/3) = 2/9$.

iv) Probability that 1 tube will be replaced after 150 hours = 2/3
Thus the probability that all the tubes will be replaced after 150 hours = $(2/3)^3 = 8/27$.

v)Suppose there are n tubes. The probability that all the n tub4es are functioning after 150 hours is$(2/3)^n$. Since this probability is given t be 0.1, we have $(2/3)^n = 0.1$. if we substitute values for n we can see that n = 5 is the maximum value of n that satisfies the equation. Hence 5 tubes are to be inserted so that all of them are functioning after 150 hours.

Example 8: The c.d.f of a continuous R.V X is given by

$F(x) = 0, x < 0$
$= x^2, 0 = x < ½$
$= 1 - 3/25(3-x)^2, ½ \leq x < 3$

Find the pdf of X and evaluate P(|X| = 1) and P(1/3 = X < 4) using both the pdf and cdf.

Solution:

The points x = 0,1/2 and 3 are points of continuity
we know that if X is a continuous random variable

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

therefore $f(x) = 0, \quad x<0$
$= 2x, \quad 0 \leq x<½$
$=6/25(3 - x), ½ \leq x<3$
$= 0, \quad x \geq 3.$

Although the points x = 1/2 , 3 are points of discontinuity for f(x), we may assume that
f(1/2) = 3/5 and f(3) = 0.

$$P(|X| \leq 1) = P(-1 \leq x \leq 1)$$
$$= \int_{-1}^{1} f(x)dx = \int_{0}^{½} 2x\,dx + \int_{½}^{1} 6/25(3 - x)dx = 13/25$$

If we use property of cdf
$P(|X| \leq 1) = P(-1 \leq x \leq 1) = F(1) - F(-1) = 13/25$

NOTES

If we use property of pdf

$$P(1/3 \le X<4) = \int_{1/3}^{1/2} 2x\,dx + \int_{1/2}^{3} 6/25(3 - x)dx = 8/9$$

If we use property of cdf

$P(1/3 \le X<4) = F(4) - F(1/3) = 1 - 1/9 = 8/9.$

Have you understood ?

Say true or false. Justify your answer.

1. A random variable is a multi valued function.
2. Probability distribution function and probability density function of a continuous random variable are continuous.
3. For a discrete random variable, the probability density function represents the probability mass function.
4. The probability mass function can take negative values.
5. A discrete random variable can be considered as a limiting case of continuous random variable with impulse distribution.

Answers:(1.False, 2. True, 3.True, 4. False, 5.True**)**

Short answer questions.

1. Define a random variable?
2. What is discrete random variable? Give an example.
3. What is continuous random variable? Give an example.
4. What is a probability distribution function?
5. Give the properties of probability distribution function?
6. What is a probability density function?
7. Give the properties of probability density function?

TRY YOURSELF !

1. A continuous random variable X has probability density function given by $f(x) = 3x^2$ $0 \le x \le 1$. Find K such that $P(X > K) = 0.05$. (Solution: k = 0.7937).
2. If the density function of a continuous RV X is given by

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

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

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

$= 0, \quad$ elsewhere

i) Find the value of a

ii) find the cdf of X

NOTES

iii)If x_1, x_2 and x_3 are 3 independent observations of X, what is the probability that exactly one of these 3 is greater than 1.5.(Solution: i)a = ½ ii) $x^4/4$, $0 \le x \le 1$; $x/2 - ¼$, $1 \le x \le 2$; $3x/2 - x^2/4 - 5/4$, $2 \le x \le 3$; $1 x > 3$. iii) 3/8)

3. A random variable X has the following probability function

X :	0	1	2	3	4
P(X):	K	3K	5K	7K	9K

i) Find the value of K

ii) Find P(X<3), P(X ≥ 3), P(0<X<4)

iii) Find the distribution function of X.

(Solution: i) K = 1/25 ii)9/25, 16/25, 3/5, iii) 1/25,4/25,9/2516/25,1)

1.5 EXPECTATION AND MOMENTS

1.5.1 Expectation

If X is a discrete random variable, then the expected value or the mean value of g(x) defined as

$$E(g(x)) = \sum_i g(x_i)\, p_i$$

Where,

$p_i = P(X = x_i)$ is the probability mass function of X.

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

$$E(g(x)) = \int_{Rx} g(x)\, f(x)\, dx$$

Mean $\mu_x = E(x) = \sum_i x_i\, p_i$ if X is discrete

$$= \int_{Rx} x\, f(x)\, dx \quad \text{if X is continuous}$$

Var(x) $= \sigma_x^2 = E((x - \mu_x)^2)$

$$= \sum (x_i - \mu_x)^2 p_i, \quad \text{if X is a discrete}$$

$$= \int (x - \mu_x)^2 f(x)dx, \quad \text{if X is continuous}$$

The square root of variance is called Standard Deviation.

NOTES

Note:
$E(a) = a$
$E(aX) = a E(X)$
$E(aX + b) = a E(X) + b$
$Var(X) = E(X^2) - (E(X))^2$
$Var(aX) = a^2 Var(X)$
$Var(aX + bY) = a^2 Var(X) + b^2 Var(Y)$

1.5.2 Moments:

If X is a discrete or continuous random variable the n^{th} moment about the origin is defined as the expected value of the n^{th} power of X and is denoted by μ_n'

$$\mu_n' = E(X^n)$$
$$= \Sigma x_i^n p_i$$
$$= \int_{-\infty}^{\infty} x^n f(x)\, dx$$

The n th central moment of a random variable X is its moment about its mean value $\bar{X}$ and is defined as

$E[(X - \bar{X})^n]$

$$= \Sigma (x_i - \bar{x})^n p_i = \mu_n \text{ , if X is discrete}$$
$$= \int_{-\infty}^{\infty} (x - \bar{x})^n f(x)\, dx = \mu_n \text{ , if X is continuous}$$

Since the first and second moments about the origin are given by $\mu_1' = E(X)$ and $\mu_2' = E(X^2)$ we have
mean = first moment about the origin

$Var(X) = \mu_2' - (\mu_1')^2$

Note:
$E(|X|^n)$ and $E(|X - \mu_x|^n)$ are called absolute moments of X.

$E\{|X - a|^n\}$ and $E\{(X - a)^n\}$ are called generalized moments of X.

1.6 MOMENT GENERATING FUNCTION:

The moment generating function (m g f) of a random variable X about the origin whose probability distribution function f(x) is given by $E(e^{tX})$ and is denoted by $M_X(t)$

Hence $M_X(t) = E(e^{tX})$

NOTES

$= \Sigma e^{tx} f(x)$, if X is a discrete

$= \int_{-\infty}^{\infty} e^{tx} f(x)\, dx$, if X is continuous

Moment generating function will exist only if the sum or integral of above definition converges.

Moment generating function generates the moments μ_n' about origin.

Consider $M_X(t) = E(e^{tX})$

$$= E\left[1 + tx + \frac{t^2X^2}{2!} + \text{- - - -} + \frac{t^nX^n}{n!} + \text{- - - -}\right]$$

$$= 1 + tE(x) + \frac{t^2E(X^2)}{2!} + \text{- - - -} + \frac{t^nE(X^n)}{n!} + \text{- - - -}$$

$$= 1 + t\,\mu_1' + \frac{t^2\,\mu_2'}{2!} + \text{- - - -} + \frac{t^n\,\mu_n'}{n!} + \text{- - - -}$$

where $\mu_n^{\,1}$ is the n^{th} moment about the origin. Thus we see that the **coefficient of $t^n / n!$** in $M_x(t)$ gives $\mu_n^{\,1}$.

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

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

$$\mu_n' = \frac{d^n}{dt^n}\left(M_X(t)\right)_{t=0}$$

Theorem: $M_{cX}(t) = M_X(ct)$ where c is a constant.

Proof:

By definition of m g f

$M_{cX}(t) = E(e^{tcX}) = E(e^{ctX}) = M_X(ct)$

Theorem: The m g f of the sum of independent random variables is equal to the product of their respective m g fs. i.e If $X_1, X_2, \text{- - - } X_n$ are n independent random variables then the m g f of $X_1 + X_2 + \text{- - - } + X_n$ is given by

$M_{X1+X2+\text{- - -}+Xn}(t) = M_{X1}(t) M_{X2}(t) \text{- - -} M_{Xn}(t)$.

Proof:

By the definition of ,m g f

$M_{X1+X2+\text{- - -}+Xn}(t) = E\left[e^{t(X1+X2+\text{- - -}+Xn)}\right]$

$= E\left[e^{tX1} e^{t\,X2} \text{- - -} e^{t\,Xn)}\right]$

NOTES

$= E[e^{tX1}][e^{tX2}] \cdots [e^{tXn})]$

$= M_{X1}(t)M_{X2}(t) - - - M_{Xn}(t).$

Theorem: (Effect of change of origin and scale on m g f):
Let X be transformed to a new variable U by changing both the origin and scale in X as $U = (X - a)/h$ where a and h are constants .Then m g f (about the origin) of U is given by
$M_U(t) = e^{-at/h} M_X(t/h)$

Proof:
$M_U(t) = E[e^{tU}] = E[e^{t(X-a)/h}]$
$= E[e^{t(X/h} e^{-at/h}]$
$= e^{-at/h} E[e^{t(X/h}]$
$= e^{-at/h} M_X(t/h)$

Note: in the above theorem putting $a = E(X) = \mu$ and $h = \sigma$ (standard deviation),
$U = [X - E(X)] / S.D(X) = (X - \mu) / \sigma = Z$ is called the standard variate.
The m g f of Z is $M_Z(t) = e^{-\mu t/h} M_X(t / \sigma)$

Theorem: Uniqueness theorem
If two random variables have the same moment generating function then they must have the same distribution.

Example1:
Find the m g f for $f(x) = \begin{cases} 2/3 & \text{at } x = 1 \\ 1/3 & \text{at } x = 2 \\ 0 & \text{otherwise} \end{cases}$
And also find μ_1' & μ_2'
Solution:
$M_X(t) = \Sigma e^{tx} f(x) = e^t(2/3) + e^{2t}(1/3)$

$\text{Mean} = \mu_1' = \frac{d}{dt}\left(M_X(t)\right)_{t=0} = \frac{d}{dt}\left(e^t(2/3) + e^{2t}(1/3)\right)_{t=0}$

$= \left(e^t(2/3) + e^{2t}(2/3)\right)_{t=0} = 4/3$

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

NOTES

Example 2: A r.v is uniformly distributed in the interval (a,b)with p d f

$f(x) = 1/(b - a)$, $a < x < b$
$= 0$ elsewhere
Find the m g f of X . Using this find the mean and variance in two different ways .

Solution:
By definition

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

$$= \int_a^b e^{tx} (1/b-a)\, dx , \quad = \frac{e^{tb} - e^{ta}}{t (b- a)}$$

$$M_X(t) = \frac{1}{t(b - a)} \left\{ 1 + tb + \frac{t^2b^2}{2!} + \frac{t^3b^3}{3!} + \ldots . - \left[1 + ta + \frac{t^2a^2}{2!} + \frac{t^3a^3}{3!} + - - \right] \right\}$$

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

$$M_X(t) = 1 + \frac{t (b + a)}{2!} + \frac{t^2(b^2 + ab + a^2)}{3!} + - - -$$

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

$$= \frac{d}{dt} \left(1 + \frac{t (b + a)}{2!} + \frac{t^2(b^2 + ab + a^2)}{3!} + - - - \right)_{t = 0}$$

$$= \left(0 + \frac{(b + a)}{2} + \text{remaining terms will contain t in the numerator} \right)$$

$$\text{Mean} = E(X) = (b+ a) / 2$$

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

$$= \frac{d^2}{dt^2} \left(1 + \frac{t (b + a)}{2!} + \frac{t^2(b^2 + ab + a^2)}{3!} + - - - \right)_{t = 0}$$

$$= 2 \frac{(b^2 + ab + a^2)}{3!}$$

NOTES

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

$= \frac{2}{3!}(b^2 + ab + a^2) - [(b+ a) / 2]^2$

$= \frac{(b - a)^2}{12}$

Another method to find mean and variance:

As you know the **coefficient of t^n / n!** in the expansion of $M_x(t)$ gives $\mu_n^{\,1}$.

Mean = $\mu_1^{\,1}$ th coefficient of t /1! or t

We have $M_X(t) = 1 + \frac{t(b + a)}{2!} + \frac{t^2(b^2 + ab + a^2)}{3!} + - - -$

$= 1 + \frac{t(b + a)}{2} + \frac{t^2(b^2 + ab + a^2)}{2! .3} + - - -$

$\mu_1^{\,1} = \frac{(b + a)}{2}$

$\mu_2 ' = \frac{(b^2 + ab + a^2)}{3}$

Variance = $\mu_2 ' - (\mu_1 ')^2 = \frac{(b - a)^2}{12}$

Example 3: Show that m g f of the random variable X having p d f

$f(x) = 1/3, -1<x<2$

$= 0,$ elsewhere is given by

$M_X(t) = \frac{e^{2t} - e^{-t}}{3t}, t \neq 0$

$= 1, \quad t = 0$

Solution:

$M_X(t) = \int_{-\infty}^{\infty} e^{tx} f(x)dx = 1/3 \int_{-1}^{2} e^{tx} dx = 1/3 [e^{tx}/t]_{-1}^{2} = 1/3[e^{2t}/t - e^{-t}/t] = \frac{e^{2t} - e^{-t}}{3t}, t \neq 0$

When t = 0 $M_X(0) = \frac{0}{0}$

Applying L'Hospital rule

$M_X(0) = \left(\frac{2e^{2t} + e^{-t}}{3}\right)_{t = 0} = 1$

NOTES

Therefore mg f of the random variable X is given by

$M_X(t) = \frac{e^{2t} - e^{-t}}{3t}$, $t \neq 0$

$= 1$, $t = 0$

Example 4:

Lt X be the random variable which assumes the value x with the probability $P(X = x) = q^{x-1}p$, x = 1,2,- —. Find the m g f of X and find its mean and variance.

Solution:

$$M_X(t) = \Sigma e^{tx} f(x) = \sum_{x=1}^{\infty} e^{tx} q^{x-1}p = \frac{p}{q}\sum_{x=1}^{\infty}(qe^t)^x = \frac{p}{q}\sum_{x=1}^{\infty}(qe^t)^{x-1}$$

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

$$M_X(t) = \frac{pe^t}{1 - qe^t}$$

$$\text{Mean} = \mu_1' = \frac{d}{dt}\left(M_X(t)\right)_{t=0} = \frac{d}{dt}\left(\frac{pe^t}{1 - qe^t}\right)_{t=0}$$

$$= \left(\frac{(1- qe^t)\, pe^t - pe^t(-qe^t)}{(1 - qe^t)^2}\right)_{t=0} = \frac{(1 - q)\, p - p(-q)}{(1 - q)^2} = \frac{p}{(1 - q)^2} = \frac{p}{p^2} = \frac{1}{p} \quad (p+ q = 1)$$

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

$$\mu_2' = \frac{d^2}{dt^2}\left(M_X(t)\right)_{t=0} = \frac{d^2}{dt^2}\left(\frac{pe^t}{1 - qe^t}\right)_{t=0} = \frac{d}{dt}\left(\frac{(1- qe^t)\, pe^t + pqe^{2t})}{(1 - qe^t)^2}\right)_{t=0}$$

$$= \left(\frac{(1-qe^t)^2\,[(1- qe^t)\, pe^t + pe^t(-qe^t)+2pqe^{2t}] - \{[(1-qe^t)\, pe^t + pqe^{2t})]2(1-qe^t)\,(-qe^t)\}}{(1 - qe^t)^4}\right)_{t=0}$$

$$= \frac{(1 - q)^2\,[(1- q)\, p + p(-q) + 2pq] - \{[(1- q)\, p + pq)]2(1 - q)\,(-q)\}}{(1 - q)^4}$$

$$\mu_2' = \frac{(1 - q)^2 p + 2pq(1 - q)}{(1 - q)^4} = \frac{(1 - q)p + 2pq}{(1 - q)^3} = \frac{p(1 + q)}{p^3} = \frac{1+q}{p^2}$$

$$\text{Var}(X) = \mu_2' - (\mu_1')^2 = 1/p - (1+q)/p^2 = \frac{q}{p^2}$$

NOTES

Example 5:

Let X have the probability mass function

$$f(x) = \begin{cases} \frac{6}{\pi^2}\frac{1}{k^2}, & k = 1, 2, \text{- - -} \\ 0 & \text{otherwise.} \end{cases}$$

Find the m g f ?

Solution:

$$M_X(t) = \Sigma e^{tx} f(x) = \sum_{k=1}^{\infty} e^{tk} \frac{6}{\pi^2}\frac{1}{k^2} \quad (\text{as } k \to 8, e^{tk} \to 8)$$

$M_X(t)$ is infinity for any $t>0$

Therefore m g f does not exist.

We will be discussing about the m g f of different distributions in the coming topics.

Have you understood ?

1. Define expectation of a random variable.
2. Define nth central moment about its origin.
3. Define mgf of a random variable?
4. State the properties of mgf of a random variable.

TRY YOURSELF !

1. Find the mgf for the distribution where
   f(x) = 2/3 at x = 1
   = 1/3 at x = 2
   = 0, otherwise
   Also find the mean and variance?
   (Solution: e^t 2/3 + e^{2t} 1/3, 4/3, 2)

2. A random variable X has density function given by
   $f(x) = 2 e^{-2x}$, $x \geq 0$
   = 0, $x < 0$
   Obtain the mgf and the mean?
   (Solution: 2/(2 - t), ½)

While constructing probabilistic models for observable phenomena, certain probability distributions arise more frequently than do others. We treat such distributions that play important roles as special probability distributions. Here we will be discussing discrete as well as continuous distributions in considerable detail.

NOTES

1.7 DISCRETE DISTRIBUTIONS

1.7.1 Binomial Distribution

Many random processes can be mapped into two outcomes. For example:

- win or lose
- success or failure
- heads or tails

These either-or situations are called binary outcomes.

Suppose I am at a party, and I ask a girl to dance. There are two outcomes. She can agree to dance with me, or she can turn me down.

At the party, I can ask any number of girls to dance with me. Each girl has the same two choices: agree to dance with me, or turn me down.

Assume that each girl's decision is independent of one another, and that my probability of success with each girl is the same. These are the assumptions that we described at the end of the last lecture.

Let n be the number of girls I ask to dance. Let X be the number of girls who agree to dance with me. We say that X is a binomial random variable, because it is the sum of binary outcomes.

Suppose that when I ask a girl to dance with me, the probability that she will agree to do so is .15. We call this p, the probability of success.

The distribution of the random variable X—my overall number of dance partners—will depend on the number of girls I ask, n, and my probability of success, p.

It is a distribution associated with repetition of independent trial of an experiment. Each trial can result in a success with probability 'p' and a failure with probability q = 1-p.

Such a trial is know as a Bernoulli trial. Some examples of Bernoulli trials are –

1) Toss of a single coin (head or tail)
2) Performance of a student in an examination (pass or fail)

A random variable X which takes only 2 values 0 and 1 with probability q and p respectively ie P (X = 1) = p, P (x = 0) = q. Then q = 1 – p is called a Bernoulli variable and X is said to have Bernoulli distribution. P is called the parameter of Bernoulli variate.

A Bernoulli random variable with parameter rp has probability mass function given by P (X –x) = p^x q^{1-x} where x = 0,1 ….

NOTES

An experiment consisting of repeated number of Bernoulli trails is called Binomial experiment. A binomial experiment must posses the following properties

1) there must be a fixed number of trials
2) all trails must have identical probabilities of success
3) the trials must be independent of each other

1.7.1.1 Binomial distribution

Let X be the number of success in a repeated independent Bernoulli trials with probability p of success for each trial .Then X is called the Binomial random variable with parameters p and n or symbolically B (n, p)

The probability mass function of a binomial random variable given by
$P (X= x) = n Cx\ p^x q^{n-x}$, where x = 0,1,2 … n where p + q = 1

Note:
1) The name binomial distribution is given since the probabilities
$nC_x\ q^{n-x}\ p^x$ (x = 0,1,2….n) are the successive terms in the expansion of the binomial expression $(p + q)^n$
2) Binomial distribution is a legitimate probability distribution since

$$\sum_{x=0}^{n} P(X = x) = \sum_{x=0}^{n} nC_x\ q^{n-x}\ p^x$$

3) If we assume that n trials constitute a set and if we consider N sets, the frequency
function of the binomial distribution is given by
$f(x) = Np(x) = N\ nCx_r p^x q^{n-x}$: x = 0,1,2, - - -n

In many scientific and engineering applications the Binomial distribution finds application.

1.7.1.2 Additive property of a Binomial distribution

If X_1 and X_2 are 2 input binomial random variable with parameter (p_1, n_1) and (p_1, n_2) then $X_1 + X_2$ is a Binomial random variable with parameters (p, $n_1 + n_2$)

In many scientific and engineering applications the Binomial distribution finds application.
1) it is used in quality control statistics to count the number of defects of an item
2) in biology to count the number of bacteria
3) in insurance problems to count the number of causalities

NOTES

1.7.1.3 Mean and variance of the binomial distribution

$$E(X) = \sum_{i}^{n} x_i p_i$$

$$= \sum_{x=0}^{n} x.\, {}^{n}C_x p^x q^{n-x}$$

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

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

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

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

$$= np. \quad (2)$$

$$E(X^2) = \sum x_i^{\,2} p_i = \sum_{0}^{n} x^2 p_x$$

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

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

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

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

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

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

$$\text{Var}(X) = E(X^2) + \{E(X)\}^2$$

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

$$= np(1-p)$$

$$= npq$$

NOTES

1.7.1.4 M G F of the binomial distribution:

$$Mx(t) = \sum_{x=0}^{n} e^{tx} p_x$$

$$= \sum_{x=0}^{n} e^{tx}\, {}^{n}C_x p^x q^{n-x}$$

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

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

1.7.1.5 Recurrence formula for the central Moments of the binomial distribution.

By definition , the k^{th} order central moment $\mu_k = E\{X - E(X)\}^k$.

For the binomial distribution B(n,p)

$$\mu_k = \sum_{x=0}^{n} (x - np)^k {}^{n}C_x p^x q^{n-x} \qquad (1)$$

By differentiating (1) with respect to p,

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

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

$$= -n\mu_{k-1} + \sum_{x=0}^{n} nC_x (r - np)^k p^{r-1} q^{n-x-1}(x - np) \qquad (\text{since } p + q = 1)$$

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

$$= -n\mu_{k-1} + \frac{1}{pq} \mu_{k+1}$$

i.e., $\mu_{k+1} = pq \left(\frac{d\mu_k}{dp} + n\mu_{k-1} \right) \qquad (2)$

using recurrence relation (2) we may compute moments of higher order, provided moments of lower order are known Putting k = 1 in (2), we get

NOTES

$$\mu_2 = pq\left[\frac{d\mu_1}{dp} + n\mu_0\right]$$

$= npq$ (since the values $\mu_0 = 1$ and $\mu_1 = 0$)

Putting k = 2 in (2), we get

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

$$= pq\frac{d}{dp}[np(1-p)]$$

$$= npq[1 - 2p] = npq(q - p)$$

Putting k = 2 in (2), we get

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

$$= npq\frac{d}{dp}\{p(1-p)(1-2p)\} + 3npq$$

$$= npq[1 - 6p + 6p^2 + 3npq]$$

$$= npq[1 - 6pq + 3npq]$$

$$= npq[1 + 3pq(n-2)]$$

Note: μ_2 is the variance, μ_3 is a measure of skewness and μ_4 is a measure of kurtosis.

1.7.1.6 Recurrence formula for Binomial distribution distribution:

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

$$P(X = x+1) = nC_{x+1} p^{x+1} q^{n-(x+1)}$$

$$\frac{P(X = x+1)}{P(X = x)} = \frac{nC_{x+1} p^{x+1} q^{n-(x+1)}}{nC_x p^x q^{n-x}}$$

$$P(X = x+1) = \frac{(n-x).\ p}{(x+1).\ q} \cdot P(X = x)$$

Example 1:

For a binomial distribution with parameter n = 5 and p, the probability of success = 0.3, find the probabilities of getting i) at least 3 success ii) at most 3 success iii)exactly 3 failures.

NOTES

Solution:

Let X denote the success.

Probability distribution is given $nCxp^xq^{n-x}$

Given n = 5. p= 0.3, q = 1 – p = 1 – 0.3 = 0.7

i)the probability of atleast 3 successes

$= P(X = 3) + P(X = 4) + P(X = 5)$

$= 5C_3(0.3)^3(0.7)^2 + 5C_4(0.3)^4(0.7)^1 + 5C_5(0.3)^5(0.7)0$

= 0 .1631

ii)The probability of atmost 3 successes

$= P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)$

$= (0.7)^5 + 5C_1(0.3)^1(0.7)^4 + 5C_2(0.3)^2(0.7)^3 + 5C_3(0.3)^3(0.7)^2$

= .9692

iii)The probability of exactly 3 failures

= the probability of exactly 2 successes

$= P(X = 2)$

$= 5C_2(0.3)^2(0.7)^3$

= .3807

Example 2 :

The mean of a binomial distribution is 5 and standard deviation is 2. Determine the distribution.

Solution:

Mean of the Binomial distribution is 5 i.e, np = 5

S D is 2. Therefore $\sqrt{}$vnpq = 2 or npq = 4

$$\frac{npq}{np} = \frac{4}{5}$$

q = 4/5, p = 1 – q = 1 – 4/5 = 1/5

We have np = 5, but p = 1/5

Therefore n = 25

Hence the Binomial distribution is $nCxp^xq^{n-x} = 25Cx(1/5)^x(4/5)^{n-x}$, x = 0,1,2, - - -25

Example 3:

The mean and variance of a Binomial variate are 8and 6.Find P(X=2).

Solution:

Given np = 8and npq = 6.

$$\frac{npq}{np} = \frac{6}{8} = \frac{3}{4}$$

Therefore q = ¾, hence p = 1 – ¾ = ¼

But np = 8 i.e, n ¼ = 8 => n = 32

The Binomial distribution is $nCxp^xq^{n-x} = 32Cx(¼)^x(¾)^{32-x}$, x = 0,1,2, - - -32

NOTES

$P(X \overset{24}{12} 2) = 1 - [p(0) + p(1)]$

$= 1 - (¾)^{32} + 32C_1(¼)(¾)^{31}$

$= 1 - (¾)^{31}[¾ + 3\ 2/4]$

$= 1 - (35/4)(¾)^{31}$

Example 4:

If on an average one vessel in every ten is wrecked, find the probability that out of five vessels expected to arrive, at least four will arrive safely.

Solution:

Let p denote the probability that a vessel will arrive safely

Then p = 9/10 and q = 1/ 10

Therefore probability for at least 4 out of 5 vessels to arrive safely

$= P(X = 4) + P(X = 5)$

$= 5C_4(9/10)^4(1/10) + (9/10)^5$

$= 0.9185$

Example 5:

In a Binomial distribution consisting of 5 independent trials, the probabilities of 1 and 2 successes are 0.4096 and 0.2048respectively. Find the parameter p of the distribution.

Solution:

The Binomial distribution is $p(x) = nCx p^x q^{n-x}$

$p(1) = 5C_1 pq^4 = 0.4096$ (1)

$p(2) = 5C_2 p^2 q^3 = 0.2048$ (2)

Dividing (2) by (1)

$$\frac{10p^2q^3}{5pq^4} = \frac{0.2048}{0.4096}$$

$$\frac{2p}{q} = \frac{1}{2}$$

$$\frac{2p}{(1-p)} = \frac{1}{2}$$

$4p = 1 - p$

$p = 1/5.$

Example 6:

If X and Y are independent Binomial variables with parameters $B_1(5, ½)$ and $B_2(7, ½)$, find $P(X + Y = 3)$

NOTES

Solution:
Since X and Y are independent binomial variables with parameters (5, ½) and (7, ½), (X + Y) is a binomial variable with parameters (12, ½)
Therefore the probability distribution of the binomial variable X + Y is given by,
$P[X + Y = x] = 12\,Cx\,(½)^x(½)^{12-x}$
$P[X + Y = 3] = 12\,C_3\,(½)^3(½)^9 = 55/704$

Example 7:

Ten coins are tossed simultaneously, find the probability of getting i) at least seven heads ii) exactly seven heads iii) at most seven heads.

Solution:
n = 10
p = probability of head in the toss of a coin = ½
q = 1 –p = 1 – ½ = ½
$p(x) = nCxp^xq^{n-x}$
$p(x) = 10Cxp^xq^{10-x}$
$p(x) = 10Cx(½)^x(½)^{10-x}$
$= 10Cx\,(1/2^{10})$

i)Probability of getting at least 7 heads
$P(x \geq 7) = P(X = 7) + P(X = 8) + P(X = 9) + P(X = 10)$
$= (½)^{10}[10C_7 + 10C_8 + 10C_9 + 10C_{10}]$

$= [\underline{120 + 45 + 10 + 1}] = \underline{11}$
$1024 \qquad 64$

ii)Probability of getting exactly 7 heads
$= P(7) = 10C_7(½)^{10} = \underline{15}$
128

iii)Probability of getting at the most 7 heads

$P(x \leq 7) = P(X = 0) + P(X = 1) + P(X = 2) + - - - - + P(X = 7)$
$= 1 – [P(X = 8) + P(X = 9) + P(X = 10)]$
$= 1 – \underline{.1.}\ [10C_8 + 10C_9 + 10C_{10}]$
2^{10}
$= 1 – 56/1024$
$= 121/128.$

Example 8:
If 10% of the screws produced by an automatic machine are defective, find the probability that out of 20 screws selected at random, there are

NOTES

i)exactly two defectives
ii)at most three defectives
iii)at least two defectives
iv)between one and three defectives(inclusive).
Find also mean and variance.

Solution:

p = probability that a screw is defective
= 0.1
q = 0.9
$p(x) = 20Cx(1/10)^x(9/10)^{20-x}$ x = 0,1,2, - - - 20
i)probability exactly two are defectives

$p(x) = 20C_2(1/10)^2(9/10)^{18}$

$= 190 \frac{.9^{18}}{10^{20}}$

ii) probability that there are at most three defectives

$= P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)$

$= (9/10)^{20} + 20C_1(1/10)(9/10)^{19} + 20C_2(1/10)^2(9/10)^{18} + 20C_3(1/10)^3(9/10)^{17}$

$= (9/10)^{17}[(9/10)^3 + 20(1/10)(9/10)^2 + 190(1/10)^2(9/10) + 1140(1/10)^3]$

$= 9^{17}/10^{20}[729 + 1620 + 1710 + 1140]$

iii)Probability that there are atleast two defectives

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

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

$= 1 - [(9/10)^{20} + 20C_1(1/10)(9/10)^{19}]$

$= 1 - (9^{19}/10^{20})(9 + 20)$

$= 1 - (9^{19}/10^{20})(29)$

iv)Probability that the number of defectives is between one and three (inclusive)

$P(1 \leq X \leq 3) = P(X = 1) + P(X = 2) + P(X = 3)$

$= 20C_1(1/10)(9/10)^{19} + 20C_2(1/10)^2(9/10)^{18} + 20C_3(1/10)^3(9/10)^{17}$

$= (9^{17}/10^{20})(4470)$

v)Mean = np = 20 .(1/10) = 2
variance = npq = 20.(1/10)(9/10) = 9/5

Example 9: The probability of a successful rocket launching is p. If launching attempts are made until 3 successful launchings have occurred, what is the probability that exactly 5 attempts are necessary? what is the probability that fewer than 5attempts will be necessary?

iii)If the launching attempts are made until 3 consecutive successful launchings occur, what are the probabilities?

NOTES

Solution:

$p(x) = nCxp^xq^{n-x}$

i)Exactly 5 attempts will be required to get 3 successes, if 2 successes occur in the first four attempts and third success in the fifth attempt.

Therefore P(exactly 5 attempts are required)

= P(2 successes in 4 attempts) x P(success in the single fifth attempt)

Now, P(2 successes in 4 attempts = $4C_2p^2q^2$ (since it follows binomial distribution)

P(success in the single fifth attempt) = p (given)

Therefore P(exactly 5 attempts are required) = $4C_2p^2q^2 \times p = 6p^3q^2$.

ii)P(fewer than 5attempts are required)

= P(exactly 3 or 4 attempts are required)

= [P(2 successes in first 2 attempts) x P(success in the 3 rd attempt)] + [P(2 successes in first 3 attempts) x P(success in the 4 th attempt]

= $2C_2p^2q^0 \times p + 3C_2p^2q^1 \times p$

= $p^3 + 3p^3q = p^3(1 + 3q)$

iii)Five attempts will be required to get 3 consecutive successes, if the first 2 attempts result in failures and the last 3 attempts result in success.

Therefore required probability = $q.q.p.p.p = q^2p^3$

Three attempts will be required to get 3 consecutive success, if each attempt result in a success.

Therefore required probability =$p.p.p = p^3$

Four attempts will be required to get 3 consecutive success, if first attempt result in a failure and the remaining attempts result in a success each.

Therefore required probability = $q.p.p.p = qp^3$

Therefore, if the launching attempts are made until 3 consecutive successful launchings occur, probability = $q^2p^3 + p^3 + qp^3 = p^3(1 + q)$

Example 10: Fit a binomial distribution for the following data:

x:	0	1	2	3	4	5	6	total
f:	5	18	28	12	7	6	4	80

Solution:

To find the binomial distribution which fits the given data, we require N, n and p. We assume N = total frequency = 80 and from the given data n = 6.

To find p :

We know mean = np. From the given data we will find the mean and then equate to np so that we can find the value of p.

NOTES

x :	0	1	2	3	4	5	6	total
f :	5	18	28	12	7	6	4	80
fx:	0	18	56	36	28	30	24	192

mean = $\Sigma fx / \Sigma f = 192/80$

= 2.4

i.e, np = 2.4, hence p = 0.4 and q = 0.6

Now we will find the theoretical frequencies which is given by Np(x)

We have $p(x) = nCxp^xq^{n-x}$

When x = 0, $80p(0) = 6C_0p^0q^6 = 3.73$

When x = 1, $80p(0) = 6C_1p^1q^5 = 14.93$

When x = 2, $80p(0) = 6C_2p^2q^4 = 24.88$

When x = 3, $80p(0) = 6C_3p^3q^3 = 22.12$

When x = 4, $80p(0) = 6C_4p^4q^2 = 11.06$

When x = 5, $80p(0) = 6C_5p^5q^1 = 2.95$

When x = 6, $80p(0) = 6C_6p^6q^0 = 0.33$

Thus we have

x:	0	1	2	3	4	5	6
f:	3.73	14.93	24.88	22.12	11.06	2.95	0.33

Converting these values into whole numbers consistent with the condition that the total frequency distribution is 80, the corresponding binomial frequency distribution is as follows

x:	0	1	2	3	4	5	6
Theoretical f:	4	15	25	22	11	3	0

Example 11: Assume that half of the population is vegetarian so that the chance of an individual being a vegetarian is ½ . Assuming that 100 investigators take samples of 10 individuals each to see whether they are vegetarians, how many investigators would you expect the reports that 3 people or less were vegetarians.

Solution

p = probability that an individual is a vegetarian = ½

q = 1 – p = ½

n = Number of individuals for each investigator = 10

N = Number of investigators

The expected number of investigators reporting x persons as vegetarians

$= N.\ n C_x\ p^x\ q^{n-x}$

$= 100.\ 10 C_x\ (½)^x . (½)^{10-x}$

$= 100.\ 10 C_x\ (½)^{10}$

NOTES

Therefore the number of investigators reporting three or less as vegetarians

$= N [p(0) + p(1) + p(2) + p(3)]$

$= 100. (½) 10 [10C_0 + 10C_1 + 10C_2 + 10C_3]$

$= (100/1024) [1 + 10 + 45 + 120]$

$= \frac{176 \times 100}{1024} = \frac{275}{16} = 17$

Example 12: In a certain town, 20% samples of the population is literate and assume that 200 investigators take samples of ten individuals to see whether they are literate. How many investigators would you expect to report that three people or less are literates in the samples.

Solution

p = probability that an individual is literate = 20% = 0.2
q = 1 – p = 0.8
n = 10
$p(x) = 10C_x (0.2)^x (0.8)^{10-x}$

Therefore number of investigators reporting 3 or less as literate

$= N [p(0) + p(1) + p(2) + p(3)]$

$= 200 [(0.8)10 + 10C_1 (0.2) (0.8)^9 + 10C_2 (0.2)^2 (0.8)^8 + 10C_3 (0.2)^3 (0.8)^7]$
$= 176$

Example 13 An irregular six faced die is thrown and the probability that in 10 throws it will give five even numbers is twice the probability that it will give four even numbers. How many times in 10,000 sets of 10 throws would you expect it to give no even numbers.

Solution

Let p = probability of getting an even number in a throw of a die.

Therefore $p(x) = 10C_x p^x q^{10-x}$
Given $p = (5) = 2 p (4)$

Therefore $10C_5 p^5 q^5 = 2.10 C_4 p^4 q^6$
$\frac{10C_5}{10C_4} . p = 2q$

NOTES

$\frac{10-5+1}{5}.p = 2q$. $\frac{nC_r}{nC_{r-1}} = \frac{n-r+1}{r}$

$6p = 10(1-p)$

$p = \frac{5}{8}$, $q = \frac{3}{8}$

Therefore $p(x) = 10C_x (5/8)^x (3/8)^{10-x}$, $x = 0,1,2, \ldots .. 10$
Number of times no even number occurs in 10,000 sets of 10 throws
$= 10,000\ p(0)$

$= 10,000\ (3/8)^{10} = 1$

1.7.2 POISSON DISTIBUTION

The Poisson distribution is used to model the number of events occurring within a given time interval.

The **Poisson distribution** is a discrete probability distribution that expresses the probability of a number of events occurring in a fixed period of time if these events occur with a known average rate, and are independent of the time since the last event. It is also called as counting random variable.

17.2.1 Poisson distribution

If X is a discrete r v that can assume the values 0,1,2 - - -such that its probability mass function is given by

$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$; $x = 0,1,2 - - -: \lambda > 0$

then is said to follow a Poisson distribution with parameter λ or symbolically X is said to follow $P(\lambda)$

1.7.2.2 Properties:

The number of outcomes occurring during a time interval is independent of the number that occurs in any other disjoint time interval. So it is memory less.
The probability that a single outcome will occur during a very short time interval is proportional to the length of the time interval. It does not depend on the number of outcomes that occur outside this time interval.
The probability that mere then one outcome will occur in such a short time interval is negligible.

1.7.2.3 Mean and variance of Poisson distribution

$E(X) = \frac{\sum x_r p_r}{x}$

NOTES

$$= \sum_{x=0}^{\infty} \frac{x\, e^{-\lambda} \lambda^x}{x!} \quad (1)$$

$$= \lambda e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^{x-1}}{(x-1)!}$$

$$= \lambda e^{-\lambda} e^{\lambda} = \lambda \quad (2)$$

$$E(X^2) = \sum_x x_x^2 p_x$$

$$= \sum_{x=0}^{\infty} \{x(x-1) + x\} \frac{e^{-\lambda} \lambda^x}{x!}$$

$$= \lambda^2 e^{-\lambda} \sum_{x=2}^{\infty} \frac{\lambda^{x-2}}{(x-2)!} + \lambda$$

$$= \lambda^2 e^{-\lambda} e^{\lambda} + \lambda = \lambda^2 + \lambda$$

$$\text{Var}(X) = E(X^2) + \{E(X)\}^2$$
$$= \lambda^2 + \lambda - \lambda^2 = \lambda$$

1.7.2.4 MGF of Poisson distribution:

$$M_X(t) = \sum_{x=0}^{\infty} e^{tx} \frac{e^{-\lambda} \lambda^x}{x!}$$

$$= \sum_{x=0}^{8} e^{-\lambda} \frac{(\lambda e^t)^x}{x!}$$

$$= e^{-\lambda} \left[1 + \frac{\lambda e^t}{1!} + \frac{(\lambda e^t)^2}{2!} + \text{- - } \right]$$

$$= \left(\begin{array}{c} -\lambda \rightarrow (\lambda e^t) \\ e \quad\quad e \end{array} \right) \quad \text{or} \quad e^{-\lambda} \exp(\lambda e^t) = \mathbf{\exp(\lambda(e^t - 1))}$$

1.7.2.5 Recurrence formula for Poisson distribution:

$$P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!} \; ; \; x = 0,1,2 \text{ - - -: } \lambda > 0$$

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

NOTES

$$\frac{P(X = x + 1)}{P(X = x)} = \frac{\lambda}{(x+1)}.$$

$$P(X = x + 1) = \frac{\lambda\, P(X = x)}{(x+1)}, \quad x>0$$

1.7.2.6 Additive property of Poisson Random Variables

If X and Y are two independent Poisson Random Variables with means λ_1 & λ_2 Respectively, then X + Y is also a Poisson Random Variable with mean $\lambda_1 + \lambda_2$.

1.7.2.7 Poisson distribution as a limiting form of Binomial Distribution

Poisson distribution as a limiting form of Binomial Distribution under the following conditions.

i) n, the number of trials is indefinitely large i.e, $n \overset{23}{11} \infty$

ii) p, the constant probability of success in each trial is very small, i.e, $p \to \infty$.

iii) $np = \lambda$ is finite or $p = \lambda/n$ and $q = 1 - \lambda/n$, where λ is a positive real number.

Proof:

If X is a binomially distributed r v with parameters n and p then,

$P(X = x) = nCx\, p^x q^{n-x}$, x = 0,1,2,- - -

$$= n\frac{(n-1)(n-2)\text{ - - - }(n-(x-1))}{x!}\, p^x(1-p)^{n-x}$$

$$= n\frac{(n-1)(n-2)\text{ - - - }(n-(x-1))}{x!}\, (\lambda/n)^x(1-(\lambda/n))^{n-x} \quad \text{(on putting } p = \lambda/n)$$

$$= \frac{\lambda^x}{x!}\left[1\,(1-1/n)(1-2/n)\text{ - - - }(1-(x-1)/n)\right](1-\lambda/n)^n.\,(1-\lambda/n)^{-x}$$

$$\lim_{\substack{n\to\infty \\ np=\lambda=\text{finite}}} P(X=x) = \frac{\lambda^x}{x!}\lim_{n\to\infty}(1-1/n)(1-2/n)\text{---}(1-(x-1)/n).\lim_{n\to\infty}(1-\lambda/n)^n.\lim_{n\to\infty}(1-\lambda/n)^{-x}$$

$$= \frac{\lambda^x e^{-\lambda}}{x!},$$ which is the probability mass function of Poisson RV

$$\left(\text{since } \lim_{n\to\infty}(1-k/n) = 1 \text{ when k is finite, } \lim_{n\to\infty}(1-\lambda/n)^n = e^{-\lambda}, \lim_{n\to\infty}(1-\lambda/n)^{-x} = 1\right)$$

Therefore we may compute binomial probabilities approximately by using the corresponding Poisson probabilities, whenever n is large and p is small.

NOTES

Some examples of Poisson variables are

i)the number trains arriving in a railway station in a given time interval.
ii)the number of alpha particles emitted by a radio active source in a given time interval
iii) the number of accidents reported in a town per day.
iv)to count the number of causalities in insurance problems.

Example1: The number of accidents in a year to taxi-drivers in a city follows a Poisson distribution with mean equal to 3. Out of 1000 taxi drivers, find approximately the number of drivers with i) no accidents in a year ii) more than 3 accidents in a year

Solution:

Here mean = $\lambda = 3$

The probability function is given by

$P(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$; $x = 0,1,2 - - -: \lambda > 0$

where is the number of accidents in a year.

i)Probability of no accidents in a year $P(0) = e^{-3} = 0.0498$

therefore out of 1000 drivers, the number of drivers with no accidents in a year is $1000 \times 0.0498 = 49.8 \overset{5}{8} 50$.

ii) Probability of more than 3 accidents in a year

= 1 - Probability of not more than 3 accidents in a year

$= 1 - P(X \overset{c}{\lambda} 3) = 1 - [P(0) + P(1) + P(2) + P(3)]$

$P(1) = e^{-3}.3$

$P(2) = e^{-3}.3^2/2! = 4.5e^{-3}$

$P(3) = e^{-3}.3^3/2! = 4.5e^{-3}$

Therefore Probability of more than 3 accidents in a year

$= 1 - [P(0) + P(1) + P(2) + P(3)]$

$= 1 - e^{-3}[1 + 3 + 4.5 + 4.5]$

$= 1 - 0.6474 = 0.3526.$

therefore out of 1000 drivers, the number of drivers with more than 3 accidents in a year

is $1000 \times 0.3526 = 352.6 \overset{5}{8} 353$.

Example 2: Fit a Poisson distribution for the following data.

x:	0	1	2	3	4	5	total
f(x):	142	156	69	27	5	1	400

Solution

To find the poisson distribution which fits the given data, we require N and λ. We assume N = total frequency = 400

To find λ :

We know mean = λ

NOTES

. From the given data we will find the mean and hence λ.

x:	0	1	2	3	4	5
f :	142	156	69	27	5	1
fx:	0	156	138	81	20	5

$$\text{mean} = \frac{\sum fx}{\sum f} = \frac{400}{400} = 1$$

$\lambda = 1$

Theoretical frequencies are given by

$N \frac{e^{-\lambda}\lambda^x}{x!}$, x = 0, 1,2,3,4,5 where N = 400

Thus we get

x:	0	1	2	3	4	5
f:	147.15	147.15	73.58	24.53	6.13	1.23

Converting these values into whole numbers consistent with the condition that the total frequency distribution is 400, the corresponding binomial frequency distribution is as follows

x:	0	1	2	3	4	5
Theoretical f:	147	147	74	25	6	1

Example 3: It is known that the probability of an item produced by a certain machine will be effective is 0.05. If the produced items are sent to the market in packets of 20,find the number of packets containing at least , exactly and atmost 2 defective items in a consignment of 1000 packets using i) binomial distribution ii) poisson approximation to binomial distribution

Solution:

Using Binomial

i) p = probability that an item is defective = 0.05
q = 0.95
n = Number of independent trials considered

$P(X = x) = nCxp^xq^{n-x}$

a) P (exactly 2 defectives) = $P(X = 2) = 20C_2p^2q^{18} = 0.1887$

If N is the number of packets, each packet containing 20 items, then the number of packets containing exactly 2 defectives is given by N x P(X = 2)

= 1000 x 0.1887 = 189, approximately

b)P(at least 2 defectives) = $P(X \geq 2)$

$= 1 - [p(0) + p(1)]$

$= 1 - [20C_0(0.05)^0(0.95)^{20} + 20C_1(0.05)^1(0.95)^{19}]$

$= 0.2641$

NOTES

If N is the number of packets, each packet containing 20 items, then the number of packets containing at least 2 defectives is given by N x $P(X \overset{24}{12} 2)$
= 1000 x 0.2641 = 264, approximately

c) $P(\text{at most 2 defectives}) = P(X \overset{c}{\lambda} 2)$
$= p(0) + p(1) + p(2)$
$= 20C_0(0.05)^0(0.95)^{20} + 20C_1(0.05)^1(0.95)^{19} + 20C_2(0.05)^2(0.95)^{18}$
$= 0.9246$

If N is the number of packets, each packet containing 20 items, then the number of packets containing atmost 2 defectives is given by N x $P(X \overset{c}{\lambda} 2)$
= 1000 x 0.9246 = 925, approximately

Using Poisson

ii) since p = 0.05 is very small and n = 20 is sufficiently large, binomial distribution may be approximated by poisson distribution with parameter

$\lambda = np = 1$

therefore $P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{e^{-1}}{1!}$

a) P (exactly 2 defectives) = $P(X \overset{24}{12} 2) = \frac{e^{-1}}{2!} = 0.1839$

If N is the number of packets, each packet containing 20 items, then the number of packets containing exactly 2 defectives is given by N x P(X = 2)
= 1000 x 0.1839 = 184, approximately

b) $P(\text{at least 2 defectives}) = P(X \overset{24}{12} 2)$
$= 1 - [p(0) + p(1)]$
$= 1 - [e^{-1} + e^{-1}]$
$= 0.2642$

If N is the number of packets, each packet containing 20 items, then the number of packets containing at least 2 defectives is given by N x $P(X \overset{24}{12} 2)$
= 1000 x 0.2642 = 264, approximately

c) $P(\text{at most 2 defectives}) = P(X \overset{c}{\lambda} 2)$
$= p(0) + p(1) + p(2)$
$= 0.9197$

If N is the number of packets, each packet containing 20 items, then the number of packets containing atmost 2 defectives is given by N x $P(X \overset{c}{\lambda} 2)$
= 1000 x 0.9197 = 920, approximately

Example 4: Assume that the number of planes crossing Indian border during war between 5p.m and 6p.m is a Poisson random variable with parameter 3 and the number between 6p.m and 7p.m is a Poisson random variable with parameter 4. If these two

NOTES

random variables are independent what is the probability that more than 5 planes cross the border between 5p.m and 7 p.m?

Solution:

Let X_1 be the number of planes crossing the border between 5p.m and 6p.m
X_2 be the number of planes crossing the border between 6p.m and 7p.m. Since X_1 and X_2 are independent Poisson random variables with parameters 3 and 4 respectively, $X_1 + X_2$ is a Poisson random variable with parameter 7.

$$P(X_1 + X_2 > 5) = 1 - P(X_1 + X_2 \leq 5)$$

$$= 1 - \sum_{x=0}^{5} \frac{e^{-7}\, 7^x}{x!}$$

$$= 1 - 0.3007 = 0.6993$$

Example 5: After correcting the proofs of the first 50 pages of a book, it is found that on the average there are 3 errors per 5 pages. Use Poisson probabilities and estimate the number of pages with 0,1,2,3 errors in the whole book of 1000 pages($e^{-6} = 0.5488$)

Solution:

λ = mean = average no : of errors per page = 3/5 = 0.6

The probability that there are x errors per page is

$$p(x) = \frac{e^{-\lambda}\, \lambda^x}{x!} = \frac{e^{-0.6}\, (0.6)^x}{x!}, \quad x = 0,1,2 - - - \text{—} -$$

i)Number of pages containing no error

N x p(0) = 1000 $e^{-0.6}$ = 1000 x .5488 ˜ 549 pages

ii)Number of pages containing one error

$$N \times p(1) = 1000\, e^{-0.6} \times \frac{0.6}{1!} = 329 \text{ pages}$$

iii)Number of pages containing 2 error

$$N \times p(1) = 1000\, e^{-0.6} \times \frac{(0.6)^2}{2!} = 1000 \times 98.7 \approx 99 \text{ pages}$$

iv)Number of pages containing 3 error

$$N \times p(1) = 1000\, e^{-0.6} \times \frac{(0.6)^3}{3!} = 20 \text{ pages}$$

Example 6: Assume that the chance of an individual coal miner being killed in a Smine accident during a year is 1/1400 . Use Poisson distribution to calculate the probability that in a mine employing 350 miners, there will be atleast one fatal accident in a year
($e^{-.25} = 0.7788$)

Solution:

$$p(x) = \frac{e^{-\lambda}\, \lambda^x}{x!}, \quad x = 0,1,2 - - - \text{—} -$$

NOTES

$p = 1/1400; n = 350$

$\lambda = np = 350/1400 = 0.25$

$p(x) = \frac{e^{-0.25}(0.25)^x}{x!}$, x = 0,1,2 – - - — -

The probability that there will be atleast one fatal accident

$= P(X=1) = 1 – P(0) = 1- e^{-0.25} = 1- 0.7788= 0.2212$

Example 7: If X and Y are independent poisson random variables, show hat the conditional distribution of X, given the value of X + Y, is a Binomial distribution.

Solution:

Let X and Y follow Poisson distributions with parameters λ_1 & λ_2 respectively.

Now

$$P[X = x / (X + Y) = n] = \frac{P[X = x \text{ and } (X + Y) = n]}{P[(X + Y) = n]} = \frac{P[X = x ; Y = n – x]}{P[(X + Y) = n]}$$

$$= \frac{P[X = x].P[X = n – x]}{P[(X + Y) = n]} \text{ (by independence of X and Y)}$$

$$= \frac{[e^{-\lambda 1}.\lambda_1^x/x!][e^{-\lambda 2}.\lambda_2^{n-x}/(n-x)!]}{e^{-(\lambda 1+\lambda 2)}.(\lambda_1+\lambda_2)^n/n!}$$

$$= \frac{n!}{x!(n-x)!}\left(\frac{\lambda_1}{(\lambda_1+\lambda_2)}\right)^r\left(\frac{\lambda_2}{(\lambda_1+\lambda_2)}\right)^{n-r}$$

$$= nC_r\, p^r q^{n-r}$$

Where $p = \left(\frac{\lambda_1}{(\lambda_1+\lambda_2)}\right)^r$ $q = \left(\frac{\lambda_2}{(\lambda_1+\lambda_2)}\right)^{n-r}$

Example 8: One-fifth of the blades produced by a blade manufacturing factory turn out to be defective. The blades are supplied in packets of 10. Use Poisson distribution to calculate the approximate number of packets containing no defective and two defective blades in a consignment of 1,00,000 packets.

Solution:

$p = (1/5)/100 = .002$

$n = 10 \quad \lambda = np = 0.2$

$p(x) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{e^{-0.2}(0.2)^x}{x!}$

i)Number of packets containing no defective

NOTES

$= Np(0) = 1,00,000e^{-0.2} = 98020$

ii) Number of packets containing one defective

$$= Np(1) = 1,00,000\frac{e^{-0.2}(0.2)}{1!} = 1960$$

iii) Number of packets containing two defectives

$$= Np(2) = \frac{1,00,000e^{-0.2}(0.2)^2}{2!} = 20$$

1.7.3 GEOMETRIC DISTRIBUTION

Suppose that I am at a party and I start asking girls to dance. Let X be the number of girls that I need to ask in order to find a partner. If the first girl accepts, then X=1. If the first girl declines but the next girl accepts, then X=2. And so on.

When X=n, it means that I failed on the first n-1 tries and succeeded on the nth try. My probability of failing on the first try is (1-p). My probabilty of failing on the first two tries is (1-p)(1-p).

My probability of failing on the first n-1 tries is $(1-p)^{n-1}$. Then, my probability of succeeding on the nth try is p. Thus, we have

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

This is known as the geometric distribution

17.3.1 Geometric distribution

Definition : Let RV X denote the number of trial of a random experiment required to obtain the first success (occurrence of an event A). Obliviously X can assume the values 1,2,3 ….

Now X= x, if and only if the first (x-1) trails result in failure (occurrence of Ä) and the x^{th} trial results in success (occurrence of A) Hence

$P (X = x) = q^{x-1} p; x = 1,2,3…., 8$

Where P (A) = p and P ($\bar{A}$) = q

If X is a discrete RV that can assume the values 1,2,3,….. 8such that its probability mass function is given by

$P (X = x) = q^{x-1} p; x = 1,2,3…., 8$where $p + q = 1$

Then X is said to follow a geometric distribution.

NOTES

Note – Geometrics distribution is legitimate probability distribution since

$$\sum_{x=1}^{\infty} P(X = x) = \sum_{x=1}^{\infty} q^{x-1} p$$

$$= p (1 + q + q^2 + \ldots. \infty)$$

$$\frac{p}{1 - q} = 1$$

1.7.3.2 Mean and variance of Geometric distribution

$$E (X) = \sum_{x} x_x P_x$$

$$= \sum_{x=1}^{\infty} x q^{x-1} p$$

$$= p (1 + 2q + 3q^2 + ... + \infty)$$

$$= p (1 – q)^{-2} = 1/p$$

$$E (X^2) = \sum_{x} x^2 p_x$$

$$= \sum_{x=1}^{\infty} x^2 q^{x-1} p$$

$$= p \sum_{x=1}^{\infty} \{ x (x + 1) – x\} q^{x-1}$$

$$= p [\{1 X^2 + 3 X^3 q + 3 X^4 q^{2 + ... + \infty}\} – \{ 1 + 2q + 3q^2 + ... + \infty\}]$$

$$= p [2 (1 – q)^{-3} - (1 – q)^{-2}]$$

$$= p \{ 2/p^3 - 1/p^2 \} = 1/p^2 (2 – p) = 1/p^2 (1 + q)$$

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

NOTES

$= 1/p^2 \ (1 + q) \ - \ 1/p^2 \quad = \ q/p^2$

Note – Sometimes the probability mass function of a geometric RV X is taken as

$P (X = x) = q^x p; x = 0,1,2,\ldots\ldots \infty$ where $p + q = 1$

If this definition is assumed then

$E (X) = q/p$ and $Var (X) = q/p^2$

1.7.3.3 MGF of Geometric distribution:

$$Mx(t) = E(e^{tx}) = \sum_{x=1}^{\infty} e^{tx} p\, q^{x-1}$$

$$= pe^t \Sigma(e^t q)^{x-1}$$

$$= pe^t(1 - qe^t)^{-1}$$

$$= \frac{pe^t}{1 - qe^t}$$

1.7.3.4 Recurrence formula for Geometric distribution

We have

$P (X = x) = q^{x-1} p$

$P (X = x + 1) = q^x p$

So $\frac{P (X = x + 1)}{P (X = x)} = q$

$P (X = x + 1) = q\, P (X = x)$

Example:1 If the probability is 0.05 that a certain kind of measuring device will show excessive drift, what is the probability that the sixth of these measuring devices tested will be the first to show excessive drift?

Solution:

If the sixth device should show excessive drift then there should be 5 failures before the sixth trial, which is geometric distribution.

$P (X = x) = q^{x-1} p; x = 1,2,3\ldots, \infty$

$P (X = 6) = q^{6-1} p$

Given $p = 0.05, q = 0.95$

$P (X = 6) = (0.95)^5 0.05 = 0.039$

NOTES

Example:2 A and B shoot independently until each has hit its own target. The probabilities of their hitting the target at each shot are 3/5 and 5/7 respectively. Find the probability that B will require more shots than A.

Solution:

Let X denote the number of trials required by A to get his first success.

$p = 3/5 \quad q = 2/5$

Then X follows a geometric distribution given by

$P(X = x) = q^{x-1} p; \; x = 1,2,3\ldots, \infty$

$= 3/5.\,(2/5)^{x-1}; \; x = 1,2,3\ldots, \infty$

Let Y denote the number of trials required by B to get his first success.

$p = 5/7 \quad q = 2/7$

Then X follows a geometric distribution given by

$P(Y = x) = q^{x-1} p; \; x = 1,2,3\ldots, \infty$

$= 5/7.\,(2/7)^{x-1}; \; x = 1,2,3\ldots, \infty$

Probability that B will require more shots than A. i.e, B requires more trials to get his first success than A requires to get his first success . i.e, probability of A getting his success in the x^{th} trial and B getting his success in $x + 1^{th}$ or $x + 2^{th}$ or - - - trial.

$$= \sum_{x=1}^{\infty} P[X = x \text{ and } Y = x + 1^{th} \text{ or } x + 2^{th} \text{ or } - - - \infty]$$

$$= \sum_{x=1}^{\infty} P[X = x].P[Y = x + 1^{th} \text{ or } x + 2^{th} \text{ or } - - - \infty]$$

$$= \sum_{x=1}^{\infty} (3/5)(2/5)^{x-1} \cdot \sum_{k=1}^{\infty} (5/7)(2/7)^{x+k-1}$$

$$= (3/7) \sum_{x=1}^{\infty} (2/5)^{x-1} \sum_{k=1}^{\infty} (2/7)^{k}(2/7)^{x-1}$$

$$= (3/7) \sum_{x=1}^{\infty} (2/5)^{x-1} (2/7)^{x-1} \cdot \sum_{k=1}^{\infty} (2/7)^{k}$$

$$= (3/7) \sum_{x=1}^{\infty} (4/35)^{x-1} \sum_{k=1}^{\infty} (2/7)^{k}$$

$$= (3/7) \sum_{x=1}^{\infty} (4/35)^{x-1} \frac{(2/7)}{1 - 2/7}$$

$$= 6/35 \sum_{x=1}^{\infty} (4/35)^{x-1} = (6/35) \frac{1}{1 - 4/35}$$

$= 6/31$

NOTES

Example:3 A die is thrown until 1 appears. Assuming that the throws are independent and the probability of getting 1 is p, find the value of p so that the probability that an odd number of throws is required is equal to 0.6. Can you find a value of p so that the probability is 0.5 that an odd number of tosses is required?

Solution:

Let X denote the number of throws required to get the first success (getting 1). Then the distribution of X is geometric

$P(X = x) = q^{x-1} p; x = 1,2,3\ldots, \infty$

$P(X = \text{an odd number}) = P(X = 1 \text{ or } 3 \text{ or } 5 - - - -)$

$$= \sum_{x=1}^{\infty} P(X = 2x - 1)$$

$$= \sum_{x=1}^{\infty} pq^{2x-2}$$

$$= (p/q^2) \sum_{x=1}^{\infty} q^{2k}$$

$$= (p/q^2)[q^2 + q^4 + - -\ ]$$

$$= (p/q^2)(q^2/1-q^2) = p/(1+q)$$

Given $1/1 + q = 0.6$

That is $1/(2-p) = 0.6 \Rightarrow p = 1/3$

If $1/1 + q = 0.5$ then $1/(2-p) = 0.5 \Rightarrow p = 0$

P = 0 is meaningless because

$$P(X = \text{an odd number}) = \sum_{x=1}^{\infty} pq^{2x-2} = 0$$

Hence the value of p cannot be found.

Example 4: Establish *memoryless* property of geometric distributions, that is , if X is a discrete random variable following a Geometric distribution,

then $P\{X>m+n / X>m\} = P\{X>n\}$, where m and n are any two positive integers. Prove the converse also, if it is true.

Since X follows geometric distribution,

$P(X = x) = q^{x-1} p; x = 1,2,3\ldots, \infty,\ p + q = 1$

$$P(X>k) = \sum_{x=k+1}^{\infty} q^{x-1} p = p(q^k + q^{k+1} + q^{k+2} + - - - + \infty)$$

$$= \frac{pq^k}{1-q} = q^k.$$

$$P\{X>m+n / X>m\} = \frac{P\{X>m+n \text{ and } X>m\}}{P\{X>m\}}$$

NOTES

$$= \frac{P\{X>m+n\}}{P\{X>m\}} = \frac{q^{m+n}}{q^{m}} = q^{n} = P(X>n)$$

The converse of the above result is also true i.e, if $P\{X > m+n / X>m\} = P\{X>n\}$, where m and n are an y two positive integers, then X follows a geometric distribution.

Since X takes the values 1,2,3, - - - -, $P\{X =1\} = 1$

Let $P(>1) = q$

Now $P\{X = (x +1)\} = P(X > x) - P(X > (x +1))$ (1)

$P\{X = (x +1)\} = 1 - P(X > (x +1))$

$P(X > x)$ $P(X > x)$

$= 1 - P\{X > (x+1)/X>x\}$

$= 1 - P\{X > 1\} = 1 - q$

$P\{X = (x +1)\} = (1 - q) P(X > x)$ (2)

$= (1 - q)[P\{ X > (x -1\} - P\{X = x\}$

(from (1), on changing x to x -1)

$= (1 - q)[P\{ X > (x -1\} - (1 - q) P(X > x-1)]$

(from (2), on changing x to x -1)

$= (1 - q)q P(X > x-1)$

$= (1 - q)q^2 P(X > x-2)$

$= (1 - q)q^{x-1} P(X > 1)$

$P\{X = (x +1)\} = (1 - q)q^{x}$

$P\{X = x \}$ $= (1 - q)q^{x-1}$ where $p = 1 - q$ and x = 1,2 -, -

That is X follows geometric distribution.

Have you understood ?

Say true or false. Justify your answer.

1.Binomial distribution id continuous .
2.For binomial distribution variance < mean.
3.Mean and variance are different for Poisson distribution.
4.Poisson distribution is a symmetrical distribution.
5.Mean is always greater or equal to the variance for a geometric distribution.
6. Geometric distribution has no memory.

(Answers: 1.False, 2.True, 3.False. 4.False,5.False, 6.True)

NOTES

Short answer questions

1. Derive the moment generating functions of all the distributions discussed above.
2. Derive the mean and variance of Poisson distribution.
3. State the additive property of Poisson distribution

Try yourself !

1. Determine the binomial distribution for which the mean is 4 and variance is 3.
(Solution: 16 Cx $(1/4)^x(3/4)^{16-x}$, x = 0,1,2,- - -)

2. A and B playas game in which their chance of winning is in the ratio 3:2. Find A's chance of winning atleast three games out of five games played.
(Solution: 0.68)

3. If X is a poisson variate such that P(x =2) = 9 P(X = 4) + 90 P(X = 6), find the Variance. (Solution: 1)

4. A manufacturer of cotter pins knows that 5% of the product is defective. If he sells cotter pins the boxes of 100 and guarantees that not more than 4 pins will be defective. What is the approximate probability that a box will fail to meet the guaranteed quality.
(Solution: P(X>4) = 0.5620)

5. If X is a geometric variate taking values 1,2, - - ∞ find P(X is odd)
(Solution: 1/(1 +q))

6. If the probability that an applicant for a driver's license will pass the road test on any given trial is 0.8, what is the probability that he will finally pass the test a) on the fourth trial b) in fewer than 4 trials?
(Solution: 0.0064, 0.9984)

1.8 CONTINUOUS DISTRIBUTION

1.8.1 UNIFORM OR RECTANGULAR DISTRBUTION

A uniform distribution is a distribution of a continuous variable in which the probability of X falling within a given interval is proportional to the size of the interval.

For example, if X is uniformly distributed between 0 and 1, then the probability that X will be between 0.3 and 0.4 is .1, because there are ten intervals of width .1 each. The probability of X falling between 0.1 and 0.25 is .15.

If X is uniformly distributed between 0 and 2, then what is the probability that X will fall between 0.3 and 0.4? Between 0.1 and 0.25?

NOTES

Uniform distributions do not occur very often in nature. However, random number generators often are built to simulate the uniform distribution.

We can use a uniform random number generator to determine the winner of a raffle. Suppose that we have 247 entries, numbered one through 247. We can choose a random number between 0 and 1, multiply it by 247, and then round it to the the nearest integer in order to pick the winner.

1.8.1.1 Uniform distribution

A continuous RV X is said to follow a uniform rectangular distribution over an interval (a,b) if its pdf is given by

$f(x) = 1/(b - a) \quad a < x < b$
$\quad\quad 0, \quad\quad$ otherwise

Here a and b (b>a) are the parameters of the distributions

.

Distribution function

$$F(x) = \int_{-\infty}^{x} f(x)dx = \int_{a}^{x} 1/(b - a)\, dx = (x - a)/(b-a)$$

$$\text{Hence } F(x) = \begin{cases} 0 ; & x < a \\ (x - a)/(b-a); & 0 < x < b \\ 1 ; & x > b \end{cases}$$

1.8.1.2 Mean and Variance

$$\text{Mean} = E(X) = \int_{-\infty}^{\infty} xf(x)dx = \int_{a}^{b} [x/(b - a)]dx$$

$$= 1/ (b-a) \left[x^2/2\right]_a^b = \frac{b^2 - a^2}{2(b - a)} = \frac{b + a}{2}$$

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)dx = \int_{a}^{b} [x^2/(b - a)]dx$$

$$= 1/ (b-a) \left[x^3/2\right]_a^b = \frac{b^3 - a^3}{3(b - a)} = \frac{b^2 + ab + a^2}{3}$$

$$\text{Varaiance} = E(X^2) - (E(X))^2 = \frac{b^2 + ab + a^2}{3} - \frac{(a^2 + 2ab + b^2)}{4} = \frac{b^2 - 2ab + a^2}{12}$$

$$= \frac{(b - a)^2}{12}$$

NOTES

Mean = (b +a)/2; Variance = $[(b - a)^2]/12$

1.8.1.3 Moment Generating function

$$Mx(t) = E(e^{tx}) = \frac{1}{(b-a)} \int_a^b e^{tx}\,dx = \frac{1}{(b-a)}\left(\frac{e^{tx}}{t}\right)_a^b = \frac{e^{bt} - e^{at}}{b - a}$$

Moments

$$\mu_r' = E(X^r) = \frac{1}{(b-a)} \int_a^b x^r\,dx = \frac{1}{(b-a)}\left(\frac{x^{r+1}}{r+1}\right)_A^b = \frac{b^{r+1} - a^{r+1}}{(r+1)(b-a)}$$

Example 1:
If X is uniformly distributed over (-3,3), find the probability that a)X<2 b)|x|<2 , c)|X - 2|<2 and find k such that P[X>k] = 1/3.

Solution:

$$f(x) = 1/(b - a) \quad a < x < b$$
$$0, \quad \text{otherwise}$$
$$f(x) = 1/6 \quad -3 < x < 3$$
$$0, \quad \text{otherwise}$$

$$\text{a) } P(X<2) = \int_{-3}^{2} 1/6\,dx = \left(x/6\right)_{-3}^{2} = 5/6$$

$$\text{b) } P(|X|<2) = P(-2 < x < 2) = \int_{-2}^{2} 1/6\,dx = \left(x/6\right)_{-2}^{2} = 4/6 = 2/3$$

c) P(|X-2|<2) = P(-2 < x-2 <2) = P(0< x< 4) = P(0 <x< 3) [since interval is (-3,3)]

$$\int_0^3 1/6\,dx = \left(x/6\right)_0^3 = 3/6 = ½$$

d) P(X>k) = 1/3

$$\Rightarrow \int_k^3 1/6\,dx = \left(x/6\right)_k^3 = (3 - k)/6 = 1/3 \text{ (given)}$$

$$\Rightarrow 3 - k = 2 \Rightarrow k = 1$$

Example 2: A passenger arrives at a bus stop at 10am knowing that the bus will arrive at some time uniformly distributed between 10am and 10.30am. What is the probability that he will have to wait longer than 10min? If at 10.15am the bus has not arrived, what is the probability that he will have to wait atleast 10 additional minutes?

Solution:

NOTES

Let X denote the waiting time.
Then pdf of X is

$f(x) = \frac{1}{30} \quad 0 < x < 30$

$= 0$ otherwise

P (he will have to wait longer than 10 min)

$= P(X > 10)$

$$= \int_{10}^{30} \frac{dx}{30} = \frac{2}{3}$$

P (he has to wait 25 min / he has already waited 15 min)

$= P (X > 25 / X > 15)$

$$= \frac{P (X > 25 \cap X > 15)}{P (X > 15)}$$

$$= \frac{P (X > 25)}{P (X > 15)}$$

$$= \frac{\int_{25}^{30} \frac{dx}{30}}{\int_{15}^{30} \frac{dx}{30}}$$

$$= \frac{5}{15} = \frac{1}{3}$$

Example 3: Show that the mgf about origin for the rectangular distribution on (-a,a) is $\frac{1}{a} \sinh at$. Also show that moments of even order are given by $\mu_{2n} = \frac{a^{2n}}{(2n + 1)}$ and all moments of odd order vanish (i.e $\mu_{2n+1} = 0$)

Solution:
Mgf about origin is given by

NOTES

$$M_X(t) = E[e^{tX}] = \frac{1}{2a}\int_{-a}^{a} e^{tx}\,dx$$

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

$$= \frac{\sin hat}{at}$$

$$= \frac{1}{at}\left[at + \frac{(at)^3}{3!} + \frac{(at)^5}{5!} + \ldots\right]$$

$$= 1 + \frac{(at)^2}{3!} + \frac{(at)^4}{5!} \ldots$$

Since there are no terms with odd powers of t in $M_X(t)$, all moments of odd order vanish

i.e $\mu_{2n+1} = 0$. in particular $\mu_1' = 0$. Thus $\mu_{r=} \mu_r^1$. Hence $\mu_{2n+1} = 0$. The moments of even
order are given by μ_{2n+1} = coefficient of $\frac{t^{2n}}{(2n)!} = \frac{a^{2n}}{2n+1}$.

Example 4: If RV has the density function f (x) prove that $y = f(x) = \int_{-\infty}^{x} f(x)\,dx$ has a
rectangular distribution over (0 , 1). If

$$f(x) = \begin{cases} \frac{x-1}{2}, & 1 \le x \le 3 \\ 0, & \text{otherwise} \end{cases}$$

Determine what interval for Y will correspond to the interval $1.1 = X = 2.9$

Solution:
The RV Y is defined as $Y = F_X(x)$ the distribution function of Y is

$$F_y(y) = P(Y \le y)$$

$$= P(F_X(x) \le y)$$

$$= P(X \le f^{-1}_x(y)$$

$$= F_x[F^{-1}_x(y)] \text{ (since } P(X \le x) = F_x(x))$$

$$= y$$

NOTES

Thus the density function of y is given by

$$F_y(y) = \frac{d}{dy}[F_Y(y)] = 1$$

The range of Y is $0 \leq y \leq 1$ since the range of Fx (x) is (0, 1). Thus Y follows a uniform distribution in (0, 1)

The distribution function of X is

$$F_x(x) = \int_1^x \frac{x-1}{2}\,dx = \frac{(x-1)^2}{4}$$

Since $Y = F_x(x)$, $Y = \frac{1}{4}(X-1)^2$

Thus when $1.1 \leq X \leq 2.9$,

$$\frac{1}{4}(1.1-1)^2 \leq Y \leq \frac{1}{4}(2.9-1)^2$$

$0.0025 \leq Y \leq 0.9025$.

Example 5: Buses arrive at a specified stop at 15 min. intervals starting at 7 a m. i.e. they arrive at 7, 7:15. 7:30,7:45and so on . If a passenger arrives at the stop between 7 and 7.30 am, find the probability that he waits
a)less than 5 min for a bus and
b)atleast 12 min for a bus.

Solution
Let X denotes the time in minutes past 7 a.m, when the passenger arrives at the stop. Then X is uniformly distributed over (0,30)

$f(x) = \frac{1}{30}$ $0 < x < 30$

$= 0$ otherwise

a) The passenger will have to wait less than 5 min. if he arrives at the stop between 7:10 and 7:15 or 7:15 and 7:30.

Therefore the required probability= P(10<x<15) + P(25<x<30)

$$= \int_{10}^{15} \frac{dx}{30} + \int_{25}^{30} \frac{dx}{30} = 1/3$$

b)The passenger will have to wait atleast 12 minutes. if he arrives between 7 and 7:03 or 7:15 and 7:18

NOTES

Therefore the required probability= $P(0<x<3) + P(15<x<18)$

$$= \int_0^3 \frac{dx}{30} + \int_{15}^{18} \frac{dx}{30} = 1/5$$

Example 6: Let two independent random variables X and Y have the geometric distribution. Show that the conditional distribution of $X/(X + Y= k)$ is uniform.

Solution:

$P(X = x) = P(Y = y) = pq^{x-1}$

$P\{X/(X + Y= k)\} = P\{X = x \text{ and } (X + Y= k)\}$

$$= \frac{P\{X = x\},P(Y= k - x)\}}{\sum_{x=1}^{k-1} P\{X = x\},P(Y= k - x)\}} = \frac{pq^{x-1}.\, pq^{k-x-1}}{\sum_{x=1}^{k-1} pq^{x-1}.\, pq^{k-x-1}}$$

$$= \frac{q^{k-2}}{\sum_{x=1}^{k-1} q^{k-2}} = 1/(k-1);\ x = 1,2,3, \ldots (k-1)$$

Thus the conditional distribution of X , given that $X + y = k$, is a discrete uniform distribution.

1.8.2 EXPONENTIAL DISTRIBUTION

1.8.2.1Exponential distribution

A continuous random variable X assuming non-negative values is said to have an exponential distribution with parameter $\lambda > 0$, if its pdf is given by

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x=0 \\ 0, & \text{otherwise} \end{cases}$$

Distribution function:

The distribution function F(x) is given by

$$F(x) = \int_0^x \lambda e^{-\lambda x}\, dx = 1 - e^{-\lambda x}$$

Therefore $F(x) = \begin{cases} 1 - e^{-\lambda x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$

NOTES

1.8.2.2 Mean and variance

$$\mu_r' = E[x^r] = \int_0^\infty x^r \lambda e^{-\lambda x}\, dx = \lambda \int_0^\infty x^r e^{-\lambda x}\, dx$$

Put $\lambda x = y$

$$\mu_r' = E[x^r] = \lambda \int_0^\infty (y/\lambda)^r e^{-\lambda x}\, (dy/\lambda)$$

$$= (1/\lambda^r)\ \sqrt{(r+1)} = r!/\lambda^r$$

Mean $= E(X) = \mu_1' = 1/\lambda$

$\mu_2' = E(X^2) = 2/\lambda^2$

$Var(X) = \mu_2 = \mu_2' - (\mu_1')^2 = 2/\lambda^2 - 1/\lambda^2 = 1/\lambda^2$

Mean $= 1/\lambda$ and variance $= 1/\lambda^2$

1.8.2.3 Moment generating function

$$Mx(t) = E[e^{tx}] = \int_0^\infty e^{tx} \lambda e^{-\lambda x}\, dx$$

$$= \lambda \left[\frac{e^{-(\lambda - t)x}}{-(\lambda - t)} \right]_0^\infty = \frac{\lambda}{\lambda - t} = \left(1 - \frac{t}{\lambda}\right)^{-1}$$

1.8.2.4 Memoryless property

If X is exponentially distributed then

$P(X>s+t/X>s) = P(X>t)$ for any $s,t>0$

Proof: $P(X>s) = \int_0^\infty \lambda e^{-\lambda x}\, dx = \left[-e^{-\lambda x}\right]_S^\infty = e^{-\lambda s}$

$$P(X>s+t/X>s) = \frac{P(X>s+t \cap X>s)}{P(X>s)} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t} = P(X>t)$$

Example 1: If X is exponentially distributed prove that the probability that X exceeds its expected value is less than ½.

Solution:

Let X be exponentially distributed with parameter λ. Then

$f(x) = \lambda e^{-\lambda x}$ $x=0$ $E(X) = 1/\lambda$

NOTES

$$P(X > 1/\lambda) = \int_{1/\lambda}^{\infty} \lambda e^{-\lambda x}\, dx = e^{-1} = 0.3679 < = ½$$

Example 2: The time in hours required to repair a machine is exponentially distributed with parameter $\lambda = ½$

i)What is the probability that the repair time exceeds 2 hours.

ii) what is the conditional probability that a repair takes atleast 10 hours given that its duration exceeds 9 hours.

Solution:

Given $\lambda = ½$

Let X denote the time to repair the machine.

The density function of X is given by

$f(x) = \lambda e^{-\lambda x}$ x=0

$= ½\, e^{-x/2}$

i) the probability that the repair time exceeds 2 hours.

$$P(X>2) = \int_{2}^{\infty} \lambda e^{-\lambda x} = \int_{2}^{\infty} ½\, e^{-x/2}\, dx = e^{-1} = 0.3679$$

ii) the conditional probability that a repair takes atleast 10 hours given that its duration exceeds 9 hours is given by

$P(X>10/X>9) = P(X>9 + 1>X>9) = P(X>1)$ (using memoryless property)

$$= \int_{1}^{\infty} ½\, e^{-x/2}\, dx = e^{-0.5} = 0.6065$$

Example 3: The daily consumption of milk in excess of 20,000 liters in a town is approximate exponentially distributed with parameter 1/3000. The town has a daily stocks of 35,000 liters. What is the probability that of 2 days selected at random the stock is insufficient for both days?

Solution:

If Y denotes daily consumption of milk then X = Y – 20000 follows an exponential distribution with parameter 1/3000. Then

$f(x) = 1/3000 \cdot e^{-x/3000}$, x>0

P(stock insufficient for one day) = P(Y>35000) = P(X + 20000 > 35000)

= P(X>15000)

$$= 1/3000 \int_{15000}^{\infty} e^{-x/3000}\, dx$$

$$= e^{-5}$$

P(Stock insufficient for 2 days) = $(e^{-5})^2\, e^{-10}$

NOTES

Example 4: The mileage which car owners get with a certain kind of radial tire is a random variable having an exponential distribution with mean 40,000 km. Find the probabilities that one of these tires will last i) at least 20,000 km and ii) at most 30,000 km.

Solution:

Let X denotes the mileage obtained with the tire.

Mean = $1/\lambda = 40{,}000 \Rightarrow \lambda = 1/40000$

The density function of X is given by

$f(x) = \lambda e^{-\lambda x}$ x $\begin{matrix}24\\12\end{matrix}$ 0

$= 1/40000 \; e^{-(1/40000)x}$

i)The probability that one of these tires will last atleast 20,000 km is given by

$$P(X=20{,}000) = \int_{20{,}000}^{\infty} 1/40000 \; e^{-(1/40000)x} \, dx$$

$$= \left(\frac{-e^{-(1/40000)x}}{40000} \right)_{20{,}000}^{\infty} = e^{-0.5}$$

ii)The probability that one of those tires will last atmost 30,000 km is given by

$$P(X=30{,}000) = \int_{0}^{30{,}000} 1/40000 \; e^{-(1/40000)x} \, dx$$

$$= \left(\frac{-e^{-(1/40000)x}}{40000} \right)_{0}^{30{,}000} = -e^{-0.75} + 1 = 0.5270$$

Example 5: A company decides to manufacture a specialized type of fuse whose lifetime is a exponential distribution. On a survey it was found that there were two processes by which the fuse may be manufactured. If a process I is used, the fuse made will have an expected length of 100hrs whereas those made by process II have an expected life length of 150 hrs. Process II is twice as costly per fuse process I which will cost Rs. 6 per fuse. Further more, if a fuse lasts for less than 200 hrs for which period it is guaranteed, a loss of Rs. 40 is assessed against the manufacturer. Which process should the company adopt?

Solution:

The density function of X is given by

$f(x) = \lambda e^{-\lambda x}$ x $\begin{matrix}24\\12\end{matrix}$ 0

NOTES

If process I is used , given Expected value as 100, i.e. $1/\lambda = 100 => \lambda = 1/100$

Then $f(x) = (1/100)e^{-(x/100)}$ $x \geq 0$

$P(X \geq 200) = \int_{200}^{\infty} (1/100)e^{-(x/100)}\,dx = e^{-2}$

$P(X<200) = 1 - P(X \geq 200) = 1 - e^{-2}$

If process II is used , given Expected value as 150, i.e. $1/\lambda = 150 => \lambda = 1/150$

Then $f(x) = (1/100)e^{-(x/100)}$ $x \geq 0$

$P(X \geq 200) = \int_{200}^{\infty} (1/150)e^{-(x/150)}\,dx = e^{-4/3}$

$P(X<200) = 1 - P(X \geq 200) = 1 - e^{-4/3}$

Let C_1 and C_2 be the costs per fuse corresponding to process I and II respectively. Then

$$C_1 = \begin{cases} 6 & X \geq 200 \\ 46 & X<200 \end{cases}$$ (loss of Rs.40 + manufacturing cost Rs. 6)

Therefore $E(C_1) = 6P(X \geq 200) + 46\,P(X < 200)$

$= 6\,e^{-2} + 46(1 - e^{-2}) = 40.5866$

Similarly,

$$C_2 = \begin{cases} 12 & X \geq 200 \\ 52 & X<200 \end{cases}$$ (loss of Rs.40 + manufacturing cost Rs. 6)

Therefore $E(C_2) = 6P(X \geq 200) + 46\,P(X < 200)$

$= 12\,e^{-4/3} + 52(1 - e^{-4/3}) = 41.456$

Since $E(C_1) < E(C_2)$ process I should be adopted.

1.8.3 Normal or Gaussian distribution

The normal distribution is almost the opposite of the uniform distribution. The uniform distribution is mathematically simple but occurs rarely in nature. The normal distribution is mathematically complex but occurs frequently in nature.

1.8.3.1 Normal distribution

A continuous random variable X is said to follow a normal distribution or Gaussian distribution with parameters μ and σ, if its probability density function is given by

NOTES

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

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

Symbolically X follows $N(\mu, \sigma)$. Some times it is also given as $N(\mu, \sigma^2)$.

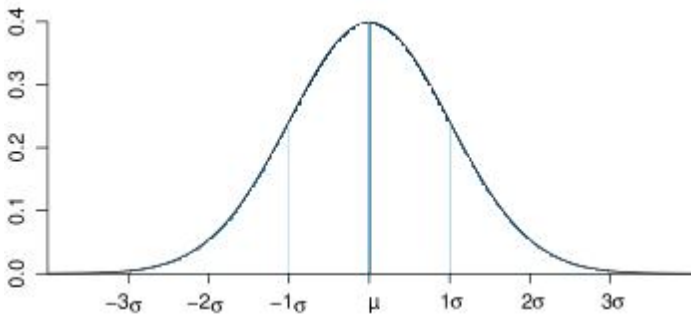
1.8.3.2 Standard Normal distribution

The normal distribution N(0, 1) is called the standardized or simply the standard normal distribution, whose density function is given by

$\Phi(z) = \frac{1}{\sqrt{2\pi}} \exp(-z^2/2)$; $-\infty < z < \infty$

This is obtained by putting $\mu = 0$ and $\sigma = 1$ and by changing x and f respectively into z and Φ. If X has distribution $N(\mu, \sigma)$ and if $Z = \frac{X - \mu}{\sigma}$, then we can prove that Z has distribution N(0,1)

1.8.3.3 Normal Probability curve



Normal probability curve

1.8.3.4 Characteristics of the Normal Distribution:

1. It is bell shaped and is symmetrical about its mean.
2. It is asymptotic to the axis, i.e., it extends indefinitely in either direction from the mean.
3. It is a continuous distribution.
4. It is a family of curves, i.e., every unique pair of mean and standard deviation defines a different normal distribution. Thus, the normal distribution is completely described by two parameters: mean and standard deviation.

NOTES

5. Total area under the curve sums to 1, i.e., the area of the distribution on each side of the mean is 0.5.
6. It is unimodal, i.e., values mound up only in the center of the curve.
7. The probability that a random variable will have a value between any two points is equal to the area under the curve between those points.

1.8.3.5 Mean and variance of Normal distribution

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

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

$$= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} (\mu + \sqrt{2}\,\sigma t) \exp(-t^2)dt \quad \left(\text{on putting } t = \frac{x-\mu}{\sigma\sqrt{2}}\right)$$

$$= \frac{\mu}{\sqrt{\pi}} \int_{-\infty}^{\infty} \exp(-t^2)dt + \frac{\sqrt{2}\,\sigma}{\sqrt{\pi}} \int_{-\infty}^{\infty} t \exp(-t^2)dt$$

(the integrand in the second part is an odd function hence it reduces to 0)

$$= \frac{\mu}{\sqrt{\pi}} \sqrt{\pi} = \mu$$

Mean $= \mu$

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

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

$$= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} (\mu + \sqrt{2}\,\sigma t)^2 \exp(-t^2)dt \quad \left(\text{on putting } t = \frac{x-\mu}{\sigma\sqrt{2}}\right)$$

$$= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \mu^2 \exp(-t^2)dt + 2\sqrt{2}\,\mu\sigma \int_{-\infty}^{\infty} t \exp(-t^2)dt + 2\sigma^2 \int_{-\infty}^{\infty} t^2 \exp(-t^2)dt$$

$$= \frac{\mu^2\sqrt{\pi}}{\sqrt{\pi}} + 0 + 2\sigma^2\, 2 \int_{0}^{\infty} t.t \exp(-t^2)dt \text{ ; (since } t^2\exp(-t^2) \text{ is an even function)}$$

$$= \mu^2 + \frac{2\sigma^2}{\sqrt{\pi}} \int_{0}^{\infty} t \exp(-t^2)\,2t\,dt$$

NOTES

Put $u = t^2$; $du = 2tdt$

$$= \mu^2 + \frac{2\sigma^2}{\sqrt{\pi}} \int_0^{\infty} u^{1/2} \exp(-u)\, du$$

$$= \mu^2 + \frac{2\sigma^2}{\sqrt{\pi}} \sqrt{(3/2)} \quad \left(\text{by definition of gamma function } \sqrt{n} = \int_0^{\infty} x^{n-1} e^{-x} dx\right)$$

$$= \mu^2 + \frac{2\sigma^2}{\sqrt{\pi}} (1/2) \,|\, (1/2) \qquad (|\, n = (n-1) \,|\, (n-1))$$

$$= \mu^2 + \frac{\sigma^2 \sqrt{\pi}}{v\pi} \qquad (|\, (1/2) = \sqrt{\pi})$$

$$= \mu^2 + \sigma^2$$

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

1 .8.3.6 Median of the normal distribution N (μ, σ)

IF X is a continuous random variable with density function f(x), then M is called the median value of X, provided that

$$\int_{-\infty}^{M} f(x)\, dx = \int_{M}^{\infty} f(x)\, dx = \tfrac{1}{2}$$

For the normal distribution N (μ, σ), the median is given by

$$\int_{M}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp(-(x-\mu)^2/2\sigma^2)\, dx = \tfrac{1}{2}$$

$$\int_{M}^{\mu} \frac{1}{\sigma\sqrt{2\pi}} \exp(-(x-\mu)^2/2\sigma^2)\, dx + \int_{\mu}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp(-(x-\mu)^2/2\sigma^2)\, dx = \tfrac{1}{2}$$

But $\int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp(-(x-\mu)^2/2\sigma^2)\, dx = 1$ and the normal curve is symmetrical about $x = \mu$

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

$$\int_{M}^{\mu} \frac{1}{\sigma\sqrt{2\pi}} \exp(-(x-\mu)^2/2\sigma^2)\, dx + \tfrac{1}{2} = \tfrac{1}{2}$$

NOTES

$$\int_{M}^{\mu} f(x)\,dx = 0$$

$M = \mu$

Therefore median = μ

1.8.3.7 Mode of the normal distribution N (μ, σ)

Mode of a continuous random variable X is defined as the values of x for which the density function f(x) is maximum.

For the normal distribution N (μ, σ)

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

$$\log f(x) = k - \frac{1}{2\sigma^2}(x-\mu)^2$$

Differentiating with respect to x,

$$\frac{f'(x)}{f(x)} = \frac{-(x-\mu)}{\sigma^2}$$

$$f'(x) = -\frac{(x-\mu) f(x)}{\sigma^2}$$

$= 0$ when $x = \mu$

$$f''(x) = \frac{-\{(x-\mu) f'(x) + f(x)\}}{\sigma^2}$$

$$\left(f''(x)\right)_{x=\mu} = -\frac{f(\mu)}{\sigma^2} < 0$$

Therefore f(x) is maximum at $x = \mu$. Therefore mode is μ

Note: Have you noticed anything while finding the mean, median and mode? Yes! For normal distribution mean, median and mode are same

1.8.3.8 Central moments of the normal distribution N (μ_r, σ)

Central moments μ_r of N (μ, σ) are given by $\mu_r = E(x-\mu)^r$

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

$$= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} (\sqrt{2}\,\sigma t)^r \exp(-t^2)\,dt$$

NOTES

$$= \frac{2^{r/2}\sigma^{r}}{v\pi} \int_{-\infty}^{\infty} t^{r} \exp(-t^2)dt$$

Case i) r is an odd integer, that is r = 2n + 1

Therefore $\mu_{2n+1} = \frac{2^{(2n+1)/2}\sigma^{2n+1}}{\sqrt{\pi}} \int_{-\infty}^{\infty} t^{2n+1}\exp(-t^2)dt$

= 0, (since the integrand is an odd function)

Case ii) r is an even integer, i.e. r = 2n

$$\mu_{2n} = \frac{2^{n}\sigma^{2n}}{\sqrt{\pi}} \int_{-\infty}^{\infty} t^{2n}\exp(-t^2)\,dt$$

$$= \frac{2^{n}\sigma^{2n}\,2}{\sqrt{\pi}} \int_{0}^{\infty} t^{2n}\exp(-t^2)\,dt \quad \text{(since } t^{2n}\exp(-t^2) \text{ is an even function of t)}$$

(put $t^2 = u$; $2tdt = du$)

$$= \frac{2^{n}\sigma^{2n}}{\sqrt{\pi}} \int_{0}^{\infty} u^{n-\frac{1}{2}} e^{-u}\,du$$

$$= \frac{2^{n}\sigma^{2n}}{\sqrt{\pi}} \lceil(n+\tfrac{1}{2}) \qquad (1)$$

$$= \frac{2^{n}\sigma^{2n}}{\sqrt{\pi}} \frac{(2n-1)}{2} \left\lceil\frac{(2n-1)}{2}\right.$$

$$= \frac{2^{n}\sigma^{2n}}{\sqrt{\pi}} \frac{(2n-1)}{2}\frac{(2n-3)}{2} \left\lceil\frac{(2n-3)}{2}\right.$$

$$= \frac{2^{n}\sigma^{2n}}{\sqrt{\pi}} \frac{(2n-1)}{2}\frac{(2n-3)}{2} \text{ - - - - } \frac{1}{2} \lceil\tfrac{1}{2}$$

(There are n terms in the series 1 . 3. 5.. . . . .(2n - 3)(2n - 1))

For example consider 1.3.5.7 we have 4 terms i.e. $\frac{7-1}{2} + 1$

i.e in general (n - 1)/2 + 1.

Therefore in the series 1 . 3. 5.. . . . .(2n - 3)(2n - 1))

there are $\frac{2n-1-1}{2} + 1 = n$)

$$= \frac{2^{n}\sigma^{2n}}{\sqrt{\pi}} \frac{1.3.5. \ldots (2n-1)}{2^{n}} \sqrt{\pi} \qquad (\lceil\tfrac{1}{2} = \sqrt{\pi})$$

$\mu_{2n} = \sigma^{2n}\, 1.3.5. \ldots (2n - 1)$

Form (1) we get

$$\mu_{2n-2} = \frac{2^{n-1}\sigma^{2n-2}}{\sqrt{\pi}} \lceil(n-\tfrac{1}{2}) \qquad (2)$$

NOTES

From (1) and (2) we get

$$\frac{\mu_{2n}}{\mu_{2n-2}} = 2\sigma^2(n - \tfrac{1}{2})$$

$$\mu_{2n} = 2\sigma^2(n - \tfrac{1}{2})\,\mu_{2n-2}$$

which gives a recurrence relation for the even order central moments of the normal distribution.

1.8.3.9 Mean deviation about the mean of the normal distribution

The central moment of the first order of a random variable X is called the mean deviation (MD) about the mean X .i.e. E{| X –E(X)| }

For the normal distribution N(μ, σ)

$$MD = \int_{-\infty}^{\infty} |x - \mu| \frac{1}{\sigma\sqrt{2\pi}} \exp(-(x - \mu)^2/2\sigma^2)\, dx$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \sqrt{2\sigma t}\, |\exp(-t^2)\, \sqrt{2}\sigma\, dt$$

$$= \frac{\sqrt{2}\,\sigma}{\sqrt{\pi}} \int_{-\infty}^{\infty} |t| \exp(-t^2)\, dt$$

$$= \frac{2\sqrt{2}\,\sigma}{\sqrt{\pi}} \int_{0}^{\infty} |t| \exp(-t^2)\, dt \quad (\text{since } |t| \exp(-t^2) \text{ is an even function. })$$

$$= \frac{\sqrt{2}\,\sigma}{\sqrt{\pi}} (\exp(-t^2))_0^{\infty}$$

$$= \frac{\sqrt{2}\,\sigma}{\sqrt{\pi}}$$

1.8.3.10 Moment generating function of N(μ, σ) and N(0,1)

The moment generating function of N(μ, σ)

$$Mx(t) = M_{\sigma Z + \mu}(t) = e^{\mu t}\, \mu_Z(\sigma t) = e^{\mu t} \exp(\sigma^2 t^2/2) = \exp[t(\mu + \sigma^2 t / 2)]$$

NOTES

Now $Mx(t) = 1 + \frac{t}{1!}(\mu + \sigma^2 t/2) + \frac{t^2}{2!}(\mu + \sigma^2 t/2)^2 + \frac{t^3}{3!}(\mu + \sigma^2 t/2)^3 + \text{- - - - -}$

$E(X)$ = Coefficient of $t/1! = \mu$

$E(X^2)$ = Coefficient of $t^2/2! = \sigma^2 + \mu^2$

$E(X^3)$ = Coefficient of $t^3/3! = 3\mu\sigma^2 + \mu^3$ and so on.

Moment generating function of N(0,1)

$Mz(t) = E(e^{tz})$

$$= \int_{-\infty}^{\infty} e^{tz}\,\varphi(z)dz$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp(-z^2/2)\, e^{tz}\, dz$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} [\exp-(z^2 - 2tz)/2]\, dz$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} [\exp-[(z-t)^2 - t^2)/2]]\, dz$$

Put $u = \frac{z-t}{\sqrt{2}}$; $du = dz/\sqrt{2}$

$$= \frac{1}{\sqrt{2\pi}} \exp(t^2/2)\sqrt{2} \int_{-\infty}^{\infty} \exp(-u^2)du$$

$$= \frac{1}{\sqrt{\pi}} \exp(t^2/2)\, |\overline{½} \qquad (\ |\overline{½} = \sqrt{\pi})$$

$$= \exp(t^2/2)$$

Note: If X has the distribution $N(\mu,\sigma)$ then $Y = aX + b$ has the distribution $N(a\mu + b, a\sigma)$ with mgf

$M_Y(t) = \exp\{[(a\mu + b)t] + [(a^2\sigma^2)t^2/2]\}$

1.8.3.11 Additive property of normal distribution

If Xi (i = 1,2 .. n) be n independent normal random variables with mean μ_i and variance σ_i^2 then

NOTES

$\sum_{i=1}^{n} a_i x_i$ is also a normal random variable with mean $\sum_{i=1}^{n} a_i \mu_i$ and variance $\sum_{i=1}^{n} a_i \sigma_i^2$.

$M_{\sum a_i x_i}(t) = M_{a_1 x_1}(t) . M_{a_2 x_2}(t) . \text{- - - - -} M_{a_n x_n}(t)$

$= \exp(a_1\mu_1 t) \exp(a_1^2\sigma_1^2 t^2/2) \times \exp(a_2\mu_2 t) \exp(a_2^2\sigma_2^2 t^2/2) \times \text{- - - -}$
$\times \exp(a_n\mu_n t) \exp(a_n^2\sigma_n^2 t^2/2) \times$

$= \exp[(\sum a_i\mu_i)t] \exp(\sum a_i^2\sigma_i^2 t^2/2)$ which is the mgf of normal random variable with mean $\sum a_i\mu_i$ and variance $\sum a_i^2\sigma_i^2$.

Note: Putting $a_1 = a_2 = 1$ and $a_3 = a_4 = \text{- - - -} = a_n = 0$, we get the following result, in particular:

If X_1 is $N(\mu_1, \sigma_1)$ and X_2 is $N(\mu_2, \sigma_2)$, then $X_1 + X_2$ is $N(\mu_1 + \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2})$.
Similarly $X_1 - X_2$ is $N(\mu_1 - \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2})$.

1.8.3.12Normal distribution as limiting form of binomial distribution.

When n is very large and neither p and q is very small, the standard normal distribution can be regarded as the limiting from of the standardised binomial . When X follows the binomial distribution B(n,p), the standrdised binomial variable Z is given by $Z = (X - np)/\sqrt{npq}$ with step size1, Z varies from $-np/\sqrt{npq}$ to $np/\sqrt{npq}$ with step size $1/\sqrt{npq}$. When neither p nor q is very small and n is very large, Z varies from - 8 to 8 with infinite small step size. Hence, in the limit, the distribution of Z may be expected to be a continuous distribution extending from - 8 to 8 , having mean 0 and SD 1.

If X follows B(n ,p), then mgf of X is given by $M_X(t) = (q + p e^t)^n$

If $Z = (X - np)/\sqrt{npq}$ then

$M_Z(t) = e^{(-npt/\sqrt{npq})} \{q + pe^{t/\sqrt{npq}}\}$

$\log M_Z(t) = \frac{-npt}{\sqrt{npq}} + n \log \{q + pe^{t/\sqrt{npq}}\}$

$= \frac{-npt}{\sqrt{npq}} + n\log\left[q + p\left\{1 + \frac{t}{\sqrt{npq}} + \frac{t^2}{2npq} + \frac{t^3}{6(npq)^{3/2}} + \text{- - - - -}\right\}\right]$

$= \frac{-npt}{\sqrt{npq}} + n\log\left[1 + \left\{\frac{pt}{\sqrt{npq}} + \frac{pt^2}{2npq} + \frac{pt^3}{6(npq)^{3/2}} + \text{- - - - -}\right\}\right]$

NOTES

$$= \frac{-npt}{\sqrt{npq}} + n\left[\frac{pt}{\sqrt{npq}}\left\{1 + \frac{t}{2\sqrt{npq}} + \frac{t^2}{6n^2p^2q^2} + \text{- - - -}\right\}\right]$$

$$\frac{-1}{2}\frac{p^2t^2}{npq}\left\{1 + \frac{t}{2\sqrt{npq}} + \frac{t^2}{6n^2p^2q^2} + \right\}^2 + \text{- -}\Bigg]$$

$= t^2/2$ + terms containing $1/\sqrt{n}$ and lower powers of n.

$$\lim_{n \to \infty} \log M_Z(t) = t^2/2$$

$$\log_e\left[\lim_{n \to \infty} M_Z(t)\right] = t^2/2$$

$$\lim_{n \to \infty} M_Z(t) = \exp(t^2/2)$$

which is the mgf of the standard normal distribution. Hence the limit of the standardized binomial distribution, as $n \to \infty$, is the standardized normal distribution.

Areas of application: Normal distribution is the most important continuous probability distribution in the field of statistics. It describes many phenomena that occur in industry , in error calculations of experiments, statistical quality control, in nature like rainfall and meteorological studies etc.

Example1: The marks obtained by a number of students in a certain subject are assumed to be approximately normally distributed with mean 55 and SD(standard deviation) 5. If 5 students are taken at random from this set, what is the probability that 3 of them would have scored marks above 60?

Solution:

If X represents the marks obtained by the students, X follows the distribution N(55,5).

P(a student scores above 60)

$= P(X > 60) = P(60 < X < \infty)$

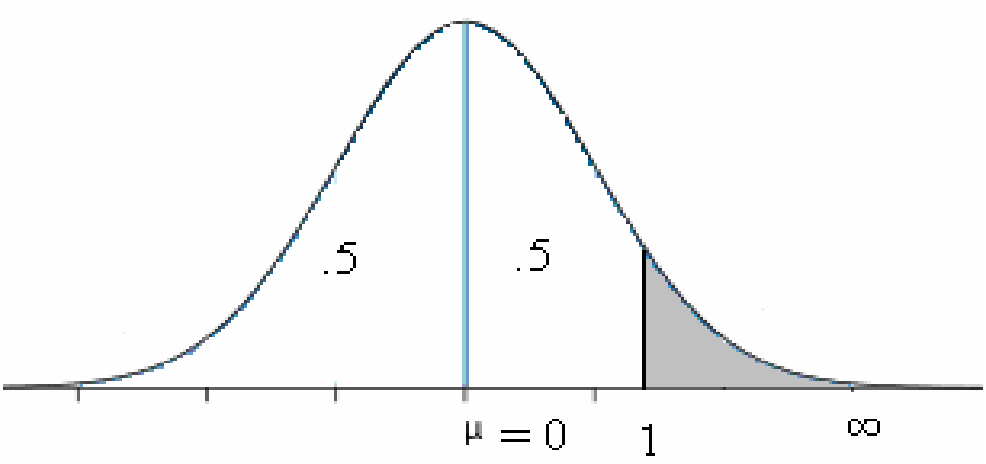
$= P[(60 - \mu)/\sigma < (X - \mu)/\sigma < \infty]$

$= P[(60 - 55)/5 < Z < \infty]$(where Z is the standard normal variate.)

NOTES

$= P[1 < Z < \infty]$

In the problem we are converting to standard normal distribution N(0,1) i.,e mean $\mu = 0$ and $\sigma = 1$.Now to find the values we will be using the normal distribution table .it will be better if you draw the normal curve to find the area. Remember the the total area under the normal curve is 1. Area to the left of $\mu = 0$ is ½ and to the right is ½.



$P[1 < Z < \infty] = 0.5 - P(0 < Z < 1)$

$=0.5 - .3413$ (from the table of areas)

$= .1587$

Therefore P(a student scores above 60) = 0.1587

Let p = P(a student scores above 60) = 0.1587

q = .8413

Since p is the same for all the students, the number Y, of students scoring above 60 follows a binomial distribution.

P(3 of the students scoring above 60 among the five students) = $nCxp^xq^{n-x}$

$= 5C_3(0.1587)^3(0.8413)^2 = 0.0283$

Example 2: The mean and SD of a certain group of 1000 high school grades, that are normally distributed are 78% and 11% respectively. Find how many grades were above 90%?

Solution:

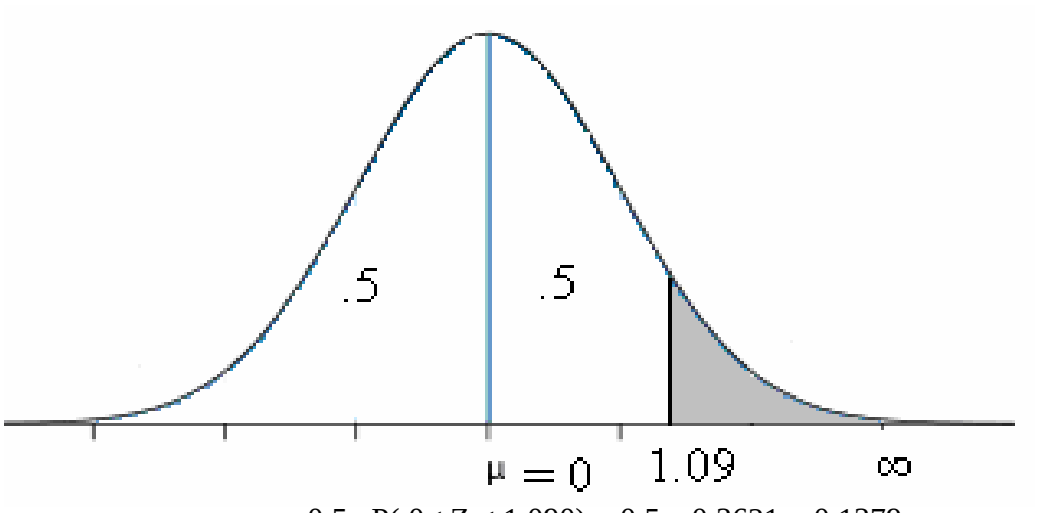
Let X represents the no: of grades. Given $\mu = 78\% = 0.78$ & $\sigma = 11\% = 0.11$

$P(X>90\%) = P(X>0.09) = P(0.09 < X < \infty])$

$= P[(0.09 - 0.78)/ 0.11 < (X - \mu)/ \sigma < \infty]$

$= P[1.090 < Z < \infty]$

NOTES



= 0.5 –P(0< Z < 1.090) = 0.5 – 0.3621 = 0.1379

Therefore out of 1000 high school grades the no : of grades above 90% will be 0.1379 x 1000 = 137.9 $\frac{5}{8}$ 138

Example 3: The local authorities in a certain city install 10,000 electric lamps in the streets of the city. If these lamps have an average life of 1000 burning hours with a SD of 200h, how many lamps might be expected to fail i) in the first 800 burning hours?ii) between 800 and 1200 burning hours. After how many burning hours would you expect iii)10% of the lamps to fail? iv) 10% of the lamps to be still burning? Assume that the life of lamps is normally distributed.

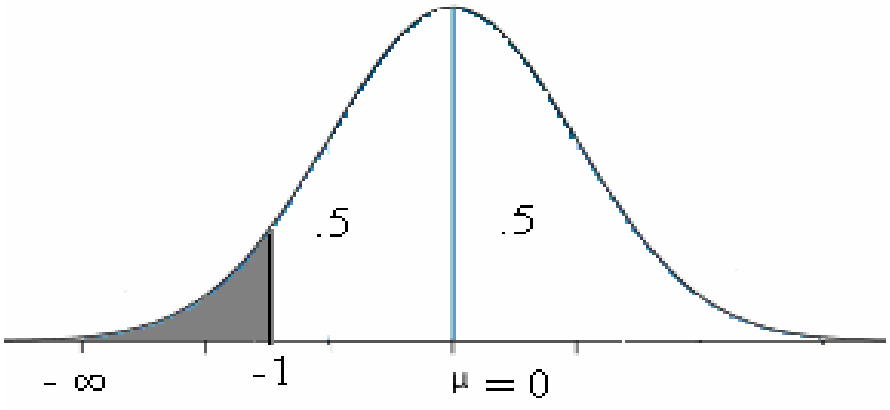
Solution:

If X represents life of the electric lamps , X follows the distribution N(1000, 200)

Total no: of electric lamps in the city, N = 10,000

i)P(X < 800) =P(-∞ <X < 800) = P(-∞ < (X – μ)/ σ < (800 – 1000)/200)

= P(-∞ < Z < -1)

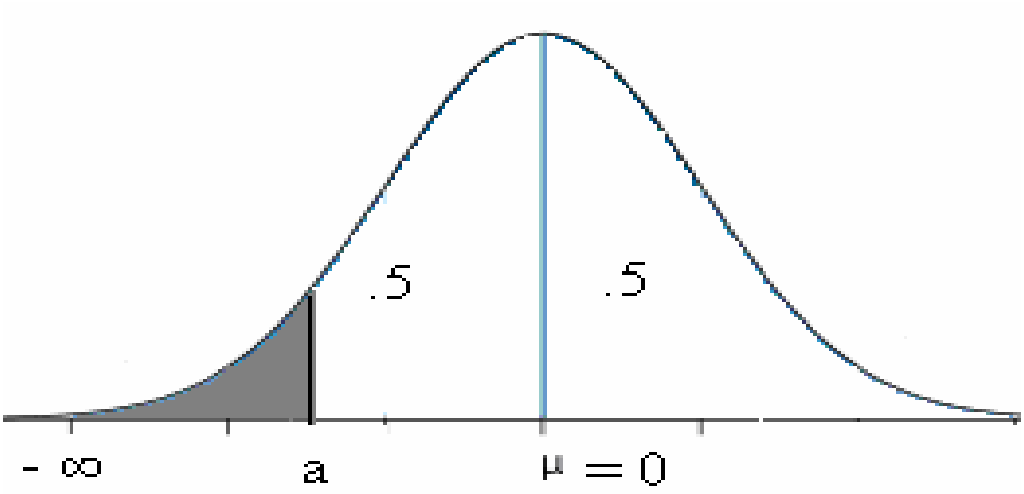


NOTES

= 0.5 - P(-1< Z< 0) = 0.5 – P(0 < Z <1) (by symmetry)
= 0.5 – 0.3413 = 0.1587
The number of lamps expected to fail in the first 800 burning hours out of 10,000 electric lamps are 10,000 x 0.1587 = 1587.

ii)P(800<x<1200) = P[(800 – 1000)/200 < (X – μ)/ σ < (1200 – 1000)/200]
= P (-1 < Z < 1) = 2 x P(0 < Z < 1) = 2 x 0.3413 = 0.6826
The number of lamps expected to fail in between 800 and 1200 burning hours out of 10,000 electric lamps are 10,000 x 0.6826 = 6826.

iii) $P(X< x_1) = 0.1$, we have to find x_1
$\Rightarrow P(-\infty< X < x_1) = P(-\infty < (X – \mu)/ \sigma < (x_1 - 1000)/200) = 0.1$
$\Rightarrow P(-\infty < Z < (x_1 - 1000)/200) = 0.1$
Let $a = (x_1 - 1000)/200)$
From $-\infty$ to a the area is given as 0.1, but we know that $-\infty$ to 0 the area is 0.5. therefore a will be lying in the left half of the normal curve.



$\Rightarrow P(-\infty < Z < (x_1 - 1000)/200) = 0.1$

$\Rightarrow 0.5 – P((x_1 - 1000)/200) < Z < 0) = 0.1$

$\Rightarrow 0.5 – P(0< Z < (1000 - x_1)/200) = 0.1$

$\Rightarrow P(0< Z < (1000 - x_1)/200)) = 0.4$ (search for the value nearer to 0.4 inside the table)

Therefore $(1000 - x_1)/200) = 1.29$

$x_1 = 744$

NOTES

iv) $P(X>x_2) = 0.1$

$\Rightarrow P(x_2 < X < 8) = 0.1$

$\Rightarrow P[(x_2 - 1000)/200 < (X - \mu)/\sigma < \infty] = 0.1$

$\Rightarrow P[(x_2 - 1000)/200 < Z < \infty] = 0.1$

Let $(x_2 - 1000)/200 = b$

$\Rightarrow 0.5 - P(0 < Z < (x_2 - 1000)/200) = 0.1$

$P(0 < Z < (x_2 - 1000)/200) = 0.4$

Therefore $(x_2 - 1000)/200) = 1.29$

$x_2 = 1256$

Example 4: The marks obtained by the students in Mathematics, Physics and Chemistry in an examination are normally distributed with the means 52, 50 and 48 and with standard deviations 10,8 and 6 respectively. Find the probability that a student selected at random has secured a total of i) 810 or above and ii) 135 or less.

Solution:

Let X,YZ Denote the marks obtained by students in mathematics, physics and chemistry respectively.

X follows N(52,10), Y follows N(50,8) and Z follows N(48,6)

By the additive property of normal distribution U = X + Y+ Z follows the distribution

$N\{52 + 50 + 48, \sqrt{10^2 + 8^2 + 6^2}\}$

i.e. N(150, 14.14)

i) $P(U=180) = P\{(180 - 150/14.14 < (U-150)/14.14 < \infty\}$

$= P\{2.12 < Z < \infty\}$

$= 0.5 - P(0 < Z < 2.12)$

$= 0.5 - 0.4830$

$= 0.0170$

ii) $P(U = 135) = P\{-\infty < U < 135\}$

$= P(-\infty < (U-150)/14.14 < (135 - 150)/14.14\}$

$= P(-\infty < Z < -1.06)$

$= 0.5 - P(-1.06 < Z < 0)$

NOTES

$= 0.5 – P(0 < Z < 1.06)$

$= 0.5 – 0.3554$

$= 0.1446$

Example 5: In a normal population with mean 15 and SD 3.5, it is found that 647 observations exceeds 16.25. What is the total number of observations in the population?

Solution:

Let N be the no: of observation.

Given that $N \times P(X > 16.25) = 647$ (1)

$P(X > 16.25) = P(16.25 < X < \infty) = P(16.25 - 15/3.5 < (X – \mu)/ \sigma < \infty)$

$= P(0.3571 < Z < \infty) = 0.5 – P(0 < Z < 0.3571)$

$= 0.5 – 0.1406 = 0.3594$

Therefore $P(X > 16.25) = 0.3594$

Substituting in (1)

$N \times 0.3594 = 647$

$N = 647 / 0.3594 = 1800$

Example 6: In a normal distribution, 7% of the items arte under 35 and 89% are under 63. what are the mean and standard deviation of the distribution? What percentage of the items are under 49?

Solution:

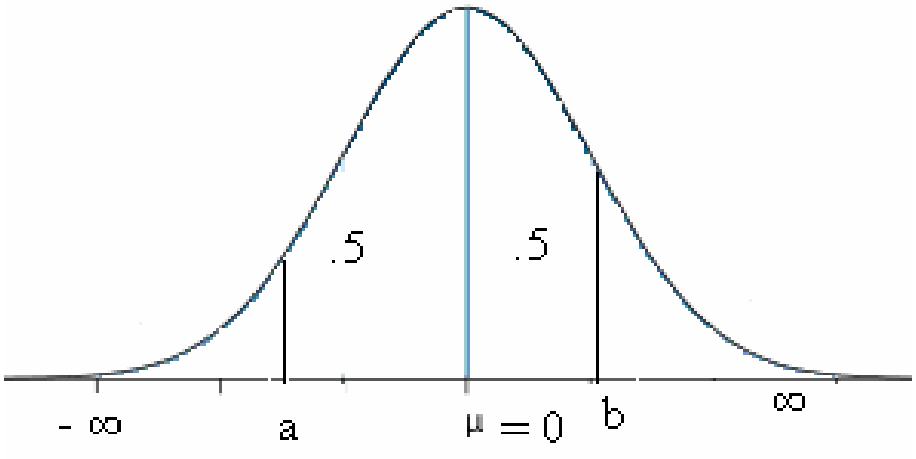
$P(X < 35) = 0.07$ $P(X < 63) = 0.89$

$P(-\infty < X < 35) = 0.07$ $P(-\infty < X < 63) = 0.89$

$P(-\infty < (X – \mu)/ \sigma < (35 - \mu)/ \sigma) = 0.07$ $P(-\infty < (X – \mu)/ \sigma < (63 – \mu)/ \sigma) = 0.89$

$P(-\infty < Z < (35 - \mu)/ \sigma) = 0.07$ $P(-\infty < Z < (63 – \mu)/ \sigma) = 0.89$

{Let $(35 - \mu)/ \sigma = a$ and $(63 – \mu)/ \sigma = b$.}



NOTES

$0.5 - P((35- \mu)/ \sigma < Z < 0) = 0.07$ $0.5 + P((63- \mu)/ \sigma < Z < 0) = 0.89$

$0.5 - P(0 < Z < (\mu -35)/ \sigma) = 0.07$ $0.5 + P((63- \mu)/ \sigma < Z < 0) = 0.89$

$P(0 < Z < (\mu -35)/ \sigma) = 0.43$ $P((63- \mu)/ \sigma < Z < 0) = 0.39$

$(\mu -35)/ \sigma = 1.47$ (1) $(63- \mu)/ \sigma = 1.23$ (2)

From 1 & 2
$35 - \mu = -1.47 \sigma$
$63 - \mu = 1.23 \sigma$

Solving these two equations we will get $\mu = 50.24$ $\sigma = 10.37$
Therefore mean = 50.24 and SD = 10.37

$P(X < 49) = P(-\infty < X < 49) = P(-\infty < (X - \mu)/ \sigma < (49 - 50.24)/10.37)$
$= P(-\infty < Z < -0.12)$
$= 0.5 - P(-0.12 < Z < 0)$
$= 0.5 - P(0 < Z < 0.12) = 0.5 - 0.0478 = 0.4522 = 45.22\%$

$P(X < 49) = 45.22\%$

Example 7: There are 400 students in the first year class of an engineering college. The probability that any student requires a copy of a particular mathematics book from the college library on any day is 0.1. How many copies of the book should be kept in t he library so that the probability may be greater than 0.95 that one of the students requiring a copy from the library has to come back disappointed?
(Use normal approximation to the binomial distribution)

Solution:
p = P(a student require book from) = 0.1
q = 0.9, n = 400

If X represents the number of students requiring the book, then X follows a binomial distribution with mean = np = 400 x 0.1 = 40 and SD = $\sqrt{npq} = 6$

In the problem it is given that X follows the distribution N(40,6)

Let x_1 be the required number of books, satisfying the given condition
i.e $P(X < x_1) > 0.95$

$P\{-\infty < X < x_1\} > 0.95$

$\Rightarrow P(-\infty < (X-40)/6 < (x_1 - 40)/6) > 0.95$

NOTES

$\Rightarrow 0.5 + P(0 < Z < (x_1 - 40)/6) > 0.95$

$\Rightarrow P(0 < Z < (x_1 - 40)/6)) > 0.45$

From the table of areas under normal curve, we find that

$P(0 < Z < 1.65) > 0.45$

Therefore $(x_1 - 40)/6 = 1.65$

- $\Rightarrow$ $m = 49.9 \overset{5}{8} 50$
- $\Rightarrow$ Therefore atleast 50 copies of the book should be kept in the library.

Example 8: If X and Y are independent random variables following N(8,2) and N(12, $4\sqrt{3}$) respectively find the value of λ such that
$P(2X - Y \overset{c}{\lambda} 2\lambda) = P(X + 2Y \overset{24}{12} \lambda)$

Solution:

Given X follows N(8,2) and Y follows N(12, $4\sqrt{3}$)

Let $A = 2X - Y$ and $B = X + 2Y$

If X_1 is $N(\mu_1, \sigma_1)$ and X_2 is $N(\mu_2, \sigma_2)$ then

$X_1 - X_2$ is $N(\mu_1 - \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2})$. (By additive property)

$2X - Y$ is $N(2x8 - 12, \sqrt{4 \text{ x } 4 + 1 \text{ x } 48})$

[since $E(aX) = a\,E(X)$, $Var(aX) = a^2 Var(X)$]]
i.e. A follows N(4,8)

$X + 2Y$ is $N(8 + 2 \text{ x } 12, \sqrt{4x4 \text{ x } 48})$
i.e B follows N(32,14)

Given $P(2X - Y \overset{c}{\lambda} 2\lambda) = P(X + 2Y \overset{24}{12} \lambda)$

$\Rightarrow P(A \overset{c}{\lambda} 2\lambda) = P(B \overset{24}{12} \lambda)$

$\Rightarrow P[(A-4)/8 \overset{c}{\lambda} (2\lambda - 4)/8] = P[(B- 32)/14 \overset{24}{12} (\lambda - 32)/14)]$

$\Rightarrow P[Z \overset{c}{\lambda} (2\lambda - 4)/8] = P[Z \overset{24}{12} (\lambda - 32)/14)]$

$\Rightarrow P[-\infty \overset{c}{\lambda} (2\lambda - 4)/8] = P[(\lambda - 32)/14) \overset{c}{\lambda} 8]$

NOTES

=> P[-∞ ≤ λ (2λ – 4)/8] = P[-∞ ≤ λ - (λ – 32)/14)] (bu symmetry)

=> (2λ – 4)/8 = - (λ – 32)/14

Therefore λ = 8.67

Normal Deviate z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2518	.2549
0.7	.2580	.2612	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4851	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4982	.4973	.4984
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4865	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
4.0	.4997									

NOTES

Have you understood ?

Say true or false. Justify your answer.

1.Exponential random variable is not memoryless.
2.Mean of uniform distribution is (b – a) /2.
3.Mean, median and mode of normal distribution are same.

(Answers:1.False,2.False,3.True**)**

Short answer questions.

1.Waht do you mean by memoryless property of exponential distribution?
2.Discuss uniform distribution and explain its mean and variance.
3.State the reproductive property of normal distribution.
4.why is normal distribution considered as an important distribution.
5.Write down the mgf of exponential distribution and hence derive its mean.

Try yourself !

1 Electric trains on a certain line run every half hour between mid-night and six in morning, What is the probability that a man entering that station at a random time during this period will have to wait atleast twenty minutes. (Solution: 1/3)

2. If X has uniform distribution in(-a, a), a>0 find a such that $P(|X| < 1) = P(|X| > 1)$ (Solution: a = 2)

3. The length of the shower in a tropical island in rainy season has an exponential distribution with parameter 2, time being measured in minutes. What is the probability that it will last for atleast one more minute? (Solution: 0.1353)

4. The marks obtained by a number of students in a certain subject are approximately normally distributed with mean 65 and standard deviation 5. If 3 students are selected at random from this group, what is the probability that atleast 1 of them would have scored above 75. (Solution: 0.0667)

5. In an examination, a student is considered to have failed, secured second class, first class and distinction, according as he scores less than 45%, between 45% and 60%, between 60% and 75% and above 75% respectively. In a particular year 10% of the students failed in examination and 5% of the students get distinction. Find the percentage of students who have got first class and second class. (Assume normal distribution of marks) (Solution: 38, 47s)

NOTES

1.9 FUNCTIONS OF RANDOM VARIABLES

1.9.1 Transformation of one dimensional random variable

Let X be a continous random variable with pdf f_x (X). We determine the density function of a new random variable Y = g (x) where g is a function. Further let y = g (x) be strictly monotonic function of x.

Now f_Y (y) the pdf of Y is given

$$f_Y(y) = f_X(x) \left| \frac{dx}{dy} \right| \text{ or } \frac{f_X(x)}{|g^1(x)|} \qquad F^1(x) = f(x)$$

g(x) is a strictly increasing function of x

case i : g (x) is a strictly increasing function of x.

F_Y (y) = P (Y $\overset{c}{\lambda}$ y), where F_Y(y) is the cdf of Y
= P[g(x) $\overset{c}{\lambda}$ y]
= P[X $\overset{c}{\lambda}$ g^{-1}(y)]
= Fx(g^{-1}(y))
Differentiating on both sides with respect toy,

$$f_Y(y) = f_X(x) \frac{dx}{dy} \text{ where } x = g^{-1}(y) \qquad (1)$$

case ii) g(x) is a strictly decreasing function of x.
F_Y (y) = P (Y $\overset{c}{\lambda}$ y), where F_Y(y) is the cdf of Y
= P[g(x) $\overset{c}{\lambda}$ y]
= P[X $\overset{24}{12}$ g^{-1}(y)]
= 1 - P[X $\overset{c}{\lambda}$ g^{-1}(y)]
= 1 - Fx(g^{-1}(y))

$$f_Y(y) = -f_X(x) \frac{dx}{dy} \qquad (2)$$

combining (1) and(2), we get

$$f_Y(y) = f_X(x) \left| \frac{dx}{dy} \right| = \frac{f_X(x)}{|g^1(x)|}$$

When $x = g^{-1}(y)$ takes finitely many values $x_1, x_2, x_3, \text{- - } x_{n, \text{then}}$

$$f_Y(y) = f_X(x_1) \frac{|dx_1|}{|dy|} + f_X(x_2) \frac{|dx_2|}{|dy|} + \text{- - - - -} + f_X(x_n) \frac{|dx_n|}{|dy|}$$

Example 1: Consider a random variable X with probability density function $f_X(x) = e^{-x}$, x=0 with the transformation $y = e^{-x}$. Find the transformed density function?

NOTES

Solution:

$f_Y(y) = f_X(x) \left| \frac{dx}{dy} \right|$

Given $f_X(x) = e^{-x}$ and $y = e^{-x}$

$\Rightarrow \frac{dy}{dx} = -e^{-x}$

Therefore $f_Y(y) = e^{-x} \left| \frac{1}{-e^{-x}} \right| = 1$

$f_Y(y) = 1, \quad 0 < y \leq 1$

$= 0, \quad$ otherwise.

Example: If X is uniformly distributed in the interval $(-\pi/2, \pi/2)$, find the pdf of $Y = \tan X$?

Solution:

As X exists in the range $(-\pi/2, \pi/2)$, $Y = \tan X$ exists in the range $-\infty$ to ∞.

As X is uniformly distributed its pdf is given by

$f(x) = 1/(b - a) \quad a < x < b$

$0, \quad$ otherwise

There $f_X(x) = 1/\pi$

$Y = \tan X \Rightarrow dy/dx = \sec^2 x$

$f_Y(y) = f_X(x) \left| \frac{dx}{dy} \right|$

$= 1/\pi \ (1 / \sec^2 x)$

$= (1/\pi)\ (1/1 + y^2)$, $-\infty < y < \infty$. [as $\sec^2 x - \tan^2 x = 1 \Rightarrow \sec^2 x = 1 + \tan^2 x$]

Example 2: The random variable Y is defined as ½ (X + |X|), where X is a random variable. Find the density and distribution function of Y ?

Solution:

Given $Y = ½ (X + |X|)$

$y = 0$ when $x < 0$

$y = x$ when $x \geq 0$

Now, distribution function

$F_Y(y) = P(Y \leq y)$

For $y < 0$, $F_Y(y) = 0$

For $y \geq 0$, $F_Y(y) = P(Y \leq y)$

$= P[(X \leq y) / X \geq 0]$

$= \frac{P(0 \leq X \leq y)}{P(X \geq 0)} = \frac{P(0 \leq X \leq y)}{1 - P(X < 0)}$

NOTES

$$= \frac{Fx(y) - Fx(0)}{1 - Fx(0)}$$

To find the density function, $F^1(x) = f(x)$

So $f_Y(y) = \frac{fx(y) - Fx(0)}{1 - Fx(0)}$

Example 3: If $Y = x^2$, where X is a Gaussian random variable with zero mean and variance σ^2, find the pdf of the random variable Y.

Solution:

$F_Y(y) = P(Y \leq y) = P(X^2 \leq y)$

$= P(-\sqrt{y} \leq X \leq \sqrt{y})$, if $y \geq 0$

$= Fx(\sqrt{y}) - Fx(-\sqrt{y})$

Differentiating with respect to y,

$f_Y(y) = \frac{1}{2\sqrt{y}}.\{f_X(\sqrt{y}) + f_X(-\sqrt{y})\}$, if $y \geq 0$ (1)

$= 0$, if $y < 0$.

It is given that X follows $N(0,\sigma)$

Therefore $f_X(x) = \frac{1}{\sigma\sqrt{2\pi}}. \exp(-x^2 / 2\sigma^2)$, $-\infty < x < \infty$

using this value in (1) , we get

$f_Y(y) = \frac{1}{\sigma\sqrt{2\pi y}}. \exp(-y / 2\sigma^2)$, $y > 0$

Example 4: Find the density function of $Y = aX + b$ in terms of the density function of X?

Solution

i)Let $a > 0$

$F_Y(y) = P(Y \leq y) = P(aX + b \leq y)$

$= P\left(X \leq \frac{(y - b)}{a}\right)$ (since $a>0$)

$= F_X\left(\frac{y - b}{a}\right)$ (1)

ii)Let $a < 0$

$F_Y(y) = P(Y \leq y) = P(aX + b \leq y)$

$= P(aX \leq y - b)$

$= P\left(X \geq \frac{y - b}{a}\right)$

$= 1 - P\left(X < \frac{(y - b)}{a}\right)$

$= 1 - F_X\left(\frac{y - b}{a}\right)$ (2)

NOTES

from (1) $f_Y(y) = \frac{1}{a} f_X\left(\frac{y-b}{a}\right)$ (3)

from (2) $f_Y(y) = -\frac{1}{a} f_X\left(\frac{y-b}{a}\right)$ (4)

combining (3) & (4)

$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right)$

Try yourself !

1.If $Y = X^2$ find the pdf of Y if X has pdf $f_X(x) = 1/3, -1 < x < 2$
= 0, elsewhere

(Solution: $f_Y(y) = 1/(3\sqrt{y}), 0 < y < 1$
$1/(6\sqrt{y}), 1 < y < 4$
0, otherwise)

2.If the density function of a continuous random variable X is given by $f_X(x) = 2/9 (x+1)$, for $-1 < x < 2$, and = 0, otherwise find the density function of $Y = ½(X + |X|)$

(Solution:
$f_Y(y) = (y+1)/4, 0<y<2$
0, otherwise)

REFERENCES:

1. T.Veerarajan, "Probability, statistics and Random Process", Tata McGraw Hill, 2002.
2. P.Kandasamy, K. Thilagavathi and K. Gunavathi, " Probability, Random Variables and Random processors", S. Chand, 2003.

NOTES

UNIT 2

TWO DIMENSIONAL RANDOM VARIABLES

- Introduction
- Two dimensional random variables
- Marginal and conditional probability
- Expectation
- Covariance
- Correlation
- Regression
- Transformation of random variables

2.1 INTRODUCTION

In the last unit we have considered one dimensional random variable. But in many practical problems several random variables interact with each other and we are interested in the joint behavior of these random variables

2.2 LEARNING OBJECTIVES

- Knowledge of basic concepts and results relating to random variables and their applications.
- Familiarity with some of the distributions commonly used to represent real-life situations.
- Knowledge of a variety of parametric and nonparametric statistical methods.
- Appreciation of the theoretical foundations of statistical methods.
- Enhancement of mathematical skills which are particularly relevant to statistics.

NOTES

2.3 TWO DIMENSIONAL RANDOM VARIABLES:

2.3.1 Definition:

Let S be the sample space of a random experiment. Let X and Y be two random variables defined on S. Then the pair (X,Y) is called a two dimensional random variable or a bivariate random variable.

2.3.2 Probability Function of (X, Y)

If (X,Y) is a two dimensional discrete random variable such that $P(X = x_i , Y = y_j) = p_{ij}$, then p_{ij} is called the probability mass function or simply the probability function of (X,Y) provided the following conditions are satisfied.

i) $p_{ij} = 0$, for all i and j

ii) $\sum_j \sum_i p_{ij} = 1$

The set of triplets $\{ x_i, y_j, p_{ij} \}$, I = 1,2,……m, …… j = 1,2,…..n,…., is called the joint probability distribution of (X,Y)

Joint Probability Density Function

If (X, Y) is a two dimensional continous random variable such that

$P\left\{ x - \frac{dx}{2} = X = x + \frac{dx}{2} \text{ and } y - \frac{dy}{2} = Y = y + \frac{dy}{2} \right\}$ = f(x,y) dx dy, then f(x,y) is called the joint pdf of (X,Y), provided f(x,y) satisfies the following conditions

i) $f(x,y) \overset{24}{12} 0$, for all $(x,y) \varepsilon R$, where R is the range space

ii) $\iint_R f(x,y)\, dx\, dy = 1$

Moreover if D is a subspace of the range space R, $P\{(X,Y) \varepsilon D\}$ is defined as $P\{(X,Y) \varepsilon D\} = \iint_D f(x,y)\, dx\, dy$. In particular

$$P\{a \overset{c}{\lambda} X \overset{c}{\lambda} b,\ c \overset{c}{\lambda} Y \overset{c}{\lambda} d\} = \int_c^d \int_a^b f(x,y)\, dx\, dy$$

2.3.3 Cumulative Distribution Function (cdf)

If (X,Y) is a two dimensional random variable (discrete or continous), then $F(x,y) = P\{X \overset{c}{\lambda} x \text{ and } Y \overset{c}{\lambda} y\}$ is called the cdf of (X,Y)

NOTES

In the discrete case,
$F(x,y) = \sum_j \sum_i p_{ij}$

In the continuous case,

$F(x,y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f(x,y)\, ds\, dy$

2.3.3.1 Properties of F (x,y)

$F(-\infty, y) = 0 = F(x, -\infty)$ and $F(\infty,\infty) = 1$
$P\{ a < X < b, Y = y \} = F(b, y) - F(a, y)$
$P\{ X \leq x, c < Y < d\} = F(x,d) - F(x,c)$
$P(a < X < b, c < Y < d\} = F(b,d) - F(a,d) - F(b,c) + F(a,c)$
At points of continuity of f(x,y)
$\frac{\partial^2 F}{\partial x \partial y} = f(x,y)$

2.4 A) MARGINAL PROBABILITY DISTRIBUTION

$P(X = x_i) = P\{ (X=x_i \text{ and } Y = y_i) \text{ or } (X = x_i \text{ and } Y = y_2) \text{ or etc } \}$

$= p_{i1} + p_{i2} + \ldots\ldots = \sum_j p_{ij}$

$P(X=x_i) = \sum_j p_{ij}$ is called the marginal probability function of X. It is defined for $X = x_1$, x_2,…… and denoted as P_{i*}. The collection of pairs $\{ x_i, p_{i*}\}$, i = 1,2,3,…..is called the marginal probability distribution of X.

Similarly the collection of pairs $\{ y_j, p_{*j}\}$, j = 1,2,3,…….is called the marginal probability distribution of Y, where $p_{*j} = \sum_i p_{ij} = P(Y = y_i)$

In the continuous case $P\{ x - \frac{1}{2} dx \leq X \leq x + \frac{1}{2} dx \}$, $-\infty < Y < \infty$

$$= \int_{-\infty}^{\infty} \int_{x-\frac{1}{2}dx}^{x+\frac{1}{2}dx} f(x,y)\, dx\, dy$$

$$= \left[\int_{-\infty}^{\infty} f(x,y)\, dy \right] dx \text{ [since f(x,y) may be treated as a constant in } (x-\tfrac{1}{2}dx, x+\tfrac{1}{2}dx)]$$

$= f_X(x)\, dx$, say

NOTES

$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy$ is called the marginal density of X.

Similarly, $f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx$ is called the marginal density of Y.

Note: $P(a \leq X \leq b) = P(a \leq X \leq b, -\infty < Y < \infty)$

$= \int_{-\infty}^{\infty}\int_{a}^{b} f(x,y)dxdy$

$= \int_{a}^{b}\int_{-\infty}^{\infty} f(x,y)dy\ dx = \int_{a}^{b} f_X(x)\,dx$

Similarly, $P(c \leq Y \leq d) = \int_{c}^{d} f_Y(y)\,dy$

2.4 B)CONDITIONAL PROBABILITY DISTRIBUTION :

$P\left[\frac{X=x_i}{Y=y_j}\right] = \frac{P\{X=x_i\ Y=y_j\}}{P\{Y=y_j\}} = \frac{p_{ij}}{p_{*j}}$ is called the conditional probability function of

X given that $Y = y_{j*}$

The collection of pairs, $\{x_i, \frac{p_{ij}}{p_{*j}}\}$ i = 1,2,3…..

is called the conditional probability distribution of X, given $Y = y_{j*}$

Similarly the collection of pairs $\{Y_j, \frac{p_{ij}}{p_{i*}}\}$ j = 1,2,3….. is called the conditional probability distribution of Y, given $X = x_{j*}$.

In the continuous case,

$P\{x - \frac{1}{2}dx \leq X < x + \frac{1}{2}dx / Y = y\}$

$= P\{x - \frac{1}{2}dx \leq X \leq x + \frac{1}{2}dx / y - \frac{1}{2}dy \leq Y \leq y + \frac{1}{2}dy\}$

$= \frac{f(x,y)\,dx\,dy}{f_Y(y)\,dy} = \{\frac{f(x,y)}{f_Y(y)}\}dx$

$\frac{f(x,y)}{f_Y(y)}$ is called the conditional density of X, given Y, and is denoted by $f_{\frac{x}{y}}$

NOTES

Similarly f(x,y) is called the conditional density of Y, given X, and is denoted by $f_{\frac{y}{x}}$
$F_X(x)$

Independent Random Variables

If (X,Y) is a two dimensional discrete random variable such that $P\{ X = x_i / Y = y_j \}$

$= P(X=x_i)$ i.e $\frac{p_{ij}}{p_{*j}} = p_{i*}$, i.e., $p_{ij} = p_{i*} \times p_{*j}$ for all i.j then X and Y are said to be independent. Similarly If (X,Y) is a two dimensional continuous random variable such that $f(x,y) = f_X(x) \times f_Y(y)$, then X and Y are said to be independent random variable.

Example 1: If X denotes the number of aces and Y the number of queens obtained when 2 cards are drawn at random **(**without replacement) from a deck of cards, obtain the joint probability distribution of (X,Y)

Solution:
Let X denote the number of aces and Y denote the number of queens. There are 4 ace cards and 4 queen cards in a deck. We are taking 2 cards from a deck of 52 cards. So X can take the values 0, 1, 2 and Y can take 0, 1, 2.

The joint probability distribution of (X,Y) is found as follows.

Y \ X	0	1	2
0	946/1326	176/1326	6/1326
1	176/1326	16/1326	0
2	6/1326	0	0

P(X = 0, Y =0) = P(drawing 2 cards none of which is a ace or queen)
= P(drawing 2 cards from the rest of 44 (52 - 8) cards)
$= \frac{44C_2}{52C_2} = \frac{946}{1326}$.

$P(X =0, Y =1) = \frac{4C_1 \times 44C_1}{52C_2} = \frac{176}{1326}$

$P(X = 0, Y = 2) = \frac{4C_2}{52C_2} = \frac{6}{1326}$.

NOTES

$P(X = 1, Y = 0) = \frac{4C_1 \times 44C_1}{52C_2} = \frac{176}{1326}$

$P(X = 1, Y = 1) = \frac{4C_1 \times 4C_1}{52C_2} = \frac{16}{1326}$

$P(X = 1, Y = 2) = 0$ (since only 2 cards are only drawn)

$P(X = 2, Y = 0) = \frac{4C_2}{52C_2} = \frac{6}{1326}$

$P(X = 2, Y = 1) = 0$

$P(X = 2, Y = 2) = 0$

Sum of all the cell probabilities = 946/1326 + 2[176/1326] + 2[6/1326] + 16/1326 = 1

Example 2 : The joint distribution of X_1 and X_2 is given by
$f(x) = \frac{x_1 + x_2}{21}$, $x_1 = 1,2$ and 3, $x_2 = 1,2$. Find a) the marginal distribution of X_1 and X_2.
b) conditional distribution of X_1 given $X_2 = 2$ and X_2 given $X_1 = 1$. Are X_1 and X_2 are independent? Also find the probability distribution of X + Y?

Solution:
The joint probability distribution of(X,Y) is given below

X_2 \ X_1	1	2	3
1	2/21	3/21	4/21
2	3/21	4/21	5/21

a)Marginal probability distribution of X_1

$X_1 = i$	$p_{i*} = \sum_{j=1}^{2} p_{ij}$	
1	$P_{11} + P_{12}$	= 2/21 + 3/21 = 5/21
2	$P_{21} + P_{22}$	= 3/21 + 4/21 = 7/21
3	P31 + P32	= 4/21 + 5/21 = 9/21

NOTES

Marginal probability distribution of X_2

$X_2 = j$	$p_{*j} = \sum_{i=1}^{3} p_{ij}$	
1	$P_{11} + P_{21} + P_{31}$	$= 2/21 + 3/21 + 4/21 = 9/21$
2	$P_{12} + P_{22} + P_{32}$	$= 3/21 + 4/21 + 5/21 = 12/21$

If X_1 and X_2 are independent we will have $p_{i*} \times p_{*j} = p_{ij}$

Let i = 2 and j = 1, $P_{2*} \times P_{*1} = 7/21 \times 9/21 = 63/21 \stackrel{?}{=} P_{21} (= 3/21)$
Therefore X_1 and X_2 are not independent.

b) Conditional distribution of X_1 given $X_2 = 2$

$X_1 = i$	$\frac{p_{i2}}{p_{*2}}$	
1	$\frac{P_{12}}{P_{12} + P_{22} + P_{32}}$	$= \frac{3/21}{3/21 + 4/21 + 5/21} = 1/4$
2	$\frac{P_{22}}{P_{12} + P_{22} + P_{32}}$	$= \frac{4/21}{3/21 + 4/21 + 5/21} = 1/3$
3	$\frac{P_{32}}{P_{12} + P_{22} + P_{32}}$	$= \frac{5/21}{3/21 + 4/21 + 5/21} = 5/12$

Conditional distribution of X_2 given $X_1 = 1$

$X_2 = j$	$\frac{p_{1j}}{p_{1*}}$	
1	$\frac{P_{11}}{P_{11} + P_{12}}$	$\frac{2/21}{2/21 + 3/21} = 2/5$
2	$\frac{P_{12}}{P_{11} + P_{12}}$	$\frac{3/21}{2/21 + 3/21} = 3/5$

NOTES

The probability distribution of X + Y is given by

X + Y	P
1	0
2	P_{11} = 2/21
3	$P_{12} + P_{21}$ = 7/21
4	$P_{22} + P_{31}$ = 8/21
5	P_{32} = 5/21

Example 3 : If the joint pdf of (X,Y) is $f(x, y) = 6\, e^{-2x-3y}$, $x \geq 0$, $y \geq 0$, find the marginal density of X and conditional density of Y given X.

Solution:

Given $f(x, y) = 6\, e^{-2x-3y}$, $x \geq 0$, $y \geq 0$

Marginal density of X is

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\, dy = \int_{0}^{\infty} 6\, e^{-2x-3y}\, dy = 6\, e^{-2x} \int_{0}^{\infty} e^{-3y}\, dy = 2e^{-2x}, \quad x \geq 0$$

Conditional density of Y given X is given by

$$f(Y/X) = \frac{f(x,y)}{f_X(x)} = \frac{6\, e^{-2x-3y}}{2e^{-2x}} = 3\, e^{-3y}, \quad y \geq 0$$

Example 4: the joint pdf of(X,Y) is given by $f(x,y) = e^{-(x+y)}$, $0 \leq x, y < 8$. Are X and Y are independent. Why?

Solution:

X and Y are independent if

$f(x,y) = f_X(x) \times f_Y(y)$

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\, dy = \int_{0}^{\infty} e^{-(x+y)}\, dy = e^{-x} \int_{0}^{\infty} e^{-y}\, dy = e^{-x}, \quad x \geq 0$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\, dy = \int_{0}^{\infty} e^{-(x+y)}\, dy = e^{-y} \int_{0}^{\infty} e^{-x}\, dy = e^{-y}, \quad y \geq 0$$

NOTES

$f_X(x) \times f_Y(y) = e^{-x} \times e^{-y} = e^{-(x+y)} = f(x,y)$.

Example 5: The joint probability mass function of (X,Y) is given by $p(x,y) = k(2x + 3y)$, $X = 0,1$; $y = 1,2,3$. Find the value of k?

Solution:
The joint probability distribution of (X,Y) is given by

X	Y		
	1	2	3
0	3k	6k	9k
1	5k	8k	11k
2	7k	10k	13k

We know that p(x, y) is a probability mass function if $\sum_j \sum_i p_{ij} = 1$

$\sum_{j=1}^{3} \sum_{i=0}^{2} p_{ij} = 1$

i.e. the sum of all the probabilities in the table is equal to 1.
i.e. $72k = 1$
therefore $k = 1/72$.

Example 6: If the joint pdf of a two –dimensional RV (X,Y) is given by

$f(x,y) = x^2 + (xy)/3,\ 0 < x < 1,\ 0 < y < 2$
$= 0$, elsewhere

i) Are X and Y independent?
ii) Find the conditional density functions? Check whether he conditional density functions are valid ?
iii) Find a)$P(X > 1/2)$ b)$P(Y < X)$ c) $P(Y < 1/2 / X < 1/2)$

Solution:

i)X and Y are independent if
$f(x,y) = f_X(x) \times f_Y(y)$

$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,dy = \int_0^2 [x^2 + (xy)/3\,dy] = 2x^2 + 2x/3,\ 0 \le x \le 1$

$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\,dx = \int_0^1 [x^2 + (xy)/3]\,dx = 1/3 + y/6,\ 0 \le y \le 2$

NOTES

$f_X(x) \times f_Y(y) = 2x^2 + 2x/3 \times 1/3 + y/6 \neq f(x,y)$

Therefore X and Y are not independent.

ii) The conditional pdf of X given Y

$$f(x/y) = \frac{f(x,y)}{f_Y(y)} = \frac{x^2 + (xy)/3}{1/3 + y/6} = \frac{6x^2 + 2xy}{2 + y} \quad 0 = x = 1, 0 = y = 2$$

The conditional pdf of Y given X

$$f(y/x) = \frac{f(x,y)}{f_X(x)} = \frac{x^2 + (xy)/3}{2x^2 + 2x/3} = \frac{3x + y}{6x + 2} \quad 0 = x = 1, 0 = y = 2$$

$$\text{Now } \int_0^1 f(x/y)\,dx = \int_0^1 \frac{6x^2 + 2xy}{2 + y}\,dx = 1$$

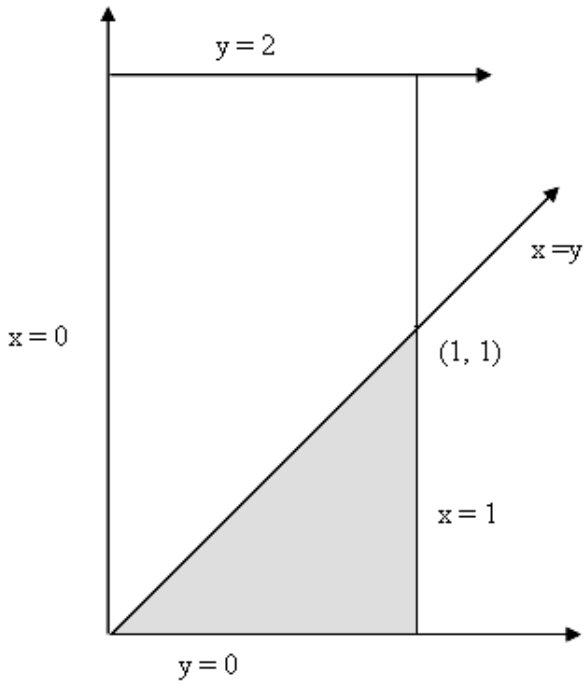
$$\text{Now } \int_0^2 f(y/x)\,dy = \int_0^2 \frac{3x + y}{6x + 2}\,dy = 1$$

So, the conditional density functions are valid.

iii) a) $P(x > ½) = \int_0^2 \int_{½}^1 [x^2 + (xy)/3]\,dx\,dy = 5/6$

b) P(Y <X)

The required region Y<X is shown in the following figure. From this figure we will find the limits for x &y.



NOTES

$$P(Y<X) = \int_0^1 \int_y^1 [x^2 + (xy)/3]\, dx\, dy = 7/24$$

c) $P(Y < 1/2 / X < 1/2) = \dfrac{P(Y < 1/2 \cap X < 1/2)}{P(X < 1/2)}$

$$P(Y < 1/2 \cap X < 1/2) = \int_0^{1/2} \int_0^{1/2} [x^2 + (xy)/3]\, dx\, dy = 5/192$$

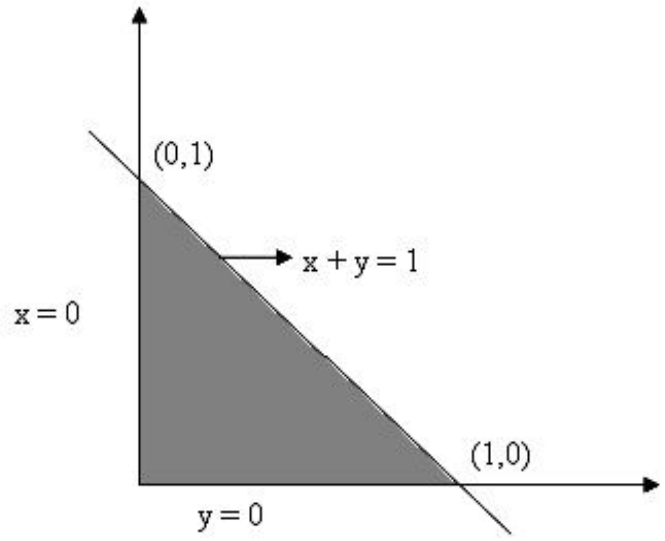
$$P(X < 1/2) = \int_0^{1/2} \int_0^{2} [x^2 + (xy)/3]\, dx\, dy = 1/6$$

Therefore $P(Y < 1/2 / X < 1/2) = \dfrac{5/192}{1/6} = 5/32$

Example 7: If $f(x, y) = e^{-(x+y)}$, $0 < x, y < \infty$
$= 0$, otherwise is a joined pdf of RV X and Y find $P(X + Y \leq 1)$

Solution:
Given $f(x, y) = e^{-(x+y)}$, $0 < x, y < \infty$
$= 0$, otherwise
To find $P(X + Y \leq 1)$

The required region $X + Y \leq 1$ is shown in the following figure. From this figure we will find the limits for x &y.



NOTES

$$P(X + Y \leq 1) = \int_0^1 \int_{\infty}^{1-y} e^{-(x+y)} dx\, dy = 1 - 2/e$$

Example 8: Given the joint pdf of two R.V. (X, Y)
$f(x,y) = kx(x - y), 0 < x < 2, |y| < x$
$= 0$, otherwise. Evaluate the value of k?

Solution:
$|y| < x \Rightarrow -x < y < x$

We know that $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y)\, dx\, dy = 1$

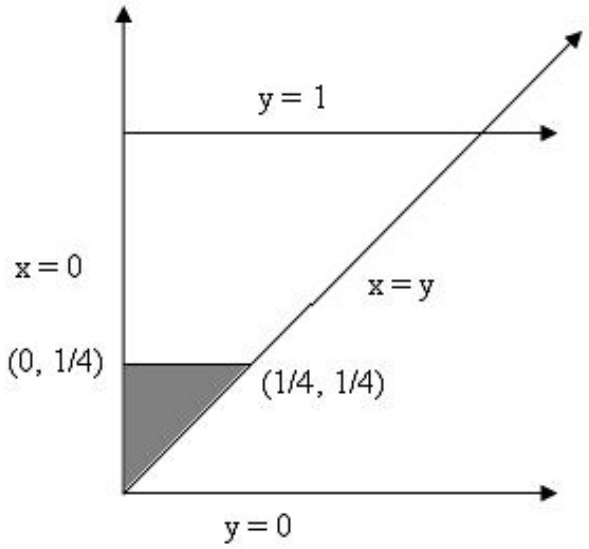
$$\int_0^2 \int_{-x}^{x} kx(x - y)\, dx\, dy = 1$$

$$8k = 1 \Rightarrow k = 1/8.$$

Example 9: The two dimensional RV (X,Y) has joint density
$f(x,y) = 8xy, 0 < x < y < 1$
$= 0$, otherwise

i) Find $P(X < \frac{1}{2} \cap Y < 1/4)$
ii) Find the marginal and conditional distributions, and
iii) Are X and Y are independent?

Solution: i)



NOTES

$$P(X < \tfrac{1}{4}, Y < 1/4) = \int_0^{1/4}\int_0^{y} 8xy\,dx\,dy = 1/256$$

ii) Marginal distribution of X

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,dy = \int_x^1 8xy\,dy = 4x - 4x^3,\ 0 < x < 1$$

Marginal distribution of Y

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\,dy = \int_0^y 8xy\,dx = 4y^3,\ 0 < y < 1$$

Conditional density of X /Y

$$f(x/y) = \frac{f(x,y)}{f_Y(y)} = \frac{8xy}{4y^3} = \frac{2x}{y^2},\ 0 < x < y < 1$$

Conditional density of Y / X

$$f(y/x) = \frac{f(x,y)}{f_X(x)} = \frac{8xy}{4x(1-x^2)} = \frac{2y}{(1-x^2)},\ 0 < x < y < 1$$

iii)To verify X, Y are independent
X and Y are independent if

$$f(x,y) = f_X(x) \times f_Y(y)$$

$$f_X(x) \times f_Y(y) = 4x - 4x^3 \times 4y^3 \neq f(x, y)$$

Therefore X and Y are not independent.

Example 10: If the joint pdf of the RV (X, Y) is given by
$f(x, y) = \frac{1}{2\pi\sigma^2} \cdot e\exp[-(x^2 + y^2)/2\sigma^2]$, $-\infty < x, y < \infty$, find $P(X^2 + Y^2 \leq a^2)$

Solution:
Her the entire xy-plane is the range space R and th vent space D is the interior of the circle $x^2 + y^2 \leq a^2$.

$$P(X^2 + Y^2 \leq a^2) = \iint_{x^2 + y^2 \leq a^2} f(x, y)dx\,dy$$

Transform from Cartesian system to polar system, i.e., put $x = r\cos\theta$ and $y = r\sin\theta$

NOTES

Then dx dy = r dr dθ
The domain of the integration becomes r = a

$$\text{Then } P(X^2 + Y^2 = a^2) = \int_0^{2\pi}\int_0^{a} \frac{1}{2\pi\sigma^2}\exp[-r^2/2\sigma^2]\, r\, dr\, d\theta$$

$$= \frac{1}{2\pi}\int_0^{2\pi} \exp[-r^2/2\sigma^2]_0^{a}\, d\theta$$

$$= 1 - \exp(-a^2/2\sigma^2)$$

Have you understood ?

1. Define joint cumulative distribution function.
2. Give the properties of cumulative distribution function.
3. Define joint probability density function.
4. Give the properties of joint probability density function
5. Define marginal probability distribution function.
6. Define marginal density function.

Try yourself !

1. Let X, Y be 2 random variables having joint pdf f(x, y) = k(x + 2y), where X, Y takes only integer values 0, 1,2. Find the marginal distribution of X & Y and conditional distribution of Y given X = 1.

Solution:

M D of X

X = i	P_{i*}
	6/27
1	9/27
2	12/27

M D of Y

Y = j	P_{*j}
0	3/27
1	9/27
2	15/27

f(X / Y = 1) = 4/9 f(Y / X = 1) = 5 /9

2. If $f(x, y) = \frac{8xy}{9}$, $1 < x < y < 2$
= 0, otherwise

NOTES

i) find the marginal density of X and Y
ii) Find he conditional density function of X, Y
iii) Verify whether X & Y are independent?

Solution:

M D of $X = \frac{4x(4 - x^2)}{9}$, $1 < x < 2$

M D of $Y = \frac{4y(y^2 - 1)}{9}$

$f(x/y) = \frac{2x}{(y^2 - 1)}$, $1 < x < y < 2$

$f(y/x) = \frac{2y}{(4 - x^2)}$, $1 < x < y < 2$

X & Y are not independent.

3. If a joint pdf $f(x,y) = \frac{(6 - x - y)}{8}$, $0 < x < 2$, $2 < y < 4$

Find i) $P(X < 1 \cap Y < 3)$
ii) $P(X + Y < 3)$
iii) $P(X < 1 / y = 3)$

Solution: $P(X < 1 \cap Y < 3) = 3/8$; $P(X + Y < 3) = 5/24$: $P(X < 1 / y = 3) = 3/5$

2.5 EXPECTATION OF A FUNCTION

If (X, Y) is a bivariate random variable and g(X,Y) is a function of X and Y, then

$E[g(x,y)] = \sum_i \sum_j g(x_i, y_j) p_{ij}$ (discrete case)

$E[g(x,y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{XY}(x,y) dx\, dy$ (continuous case)

Note:

$E(X) = \sum_i x_i p_{i*}$; $E(Y) = \sum_j y_j p_{*j}$ (discrete case)

$E(X) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f(x,y) dx dy$; $E(Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f(x,y) dx dy$ (continuous case)

$E(X + Y) = \sum_i \sum_j (x_i + y_j) p(x_i, y_j)$ (discrete case)

$E(X + Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x + y) f_{XY}(x,y)\, dx\, dy$ (continuous case)

NOTES

$E(XY) = \sum_i \sum_j (x_i y_j) p(x_i, y_j)$ (discrete case)

$E(XY) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (xy)\, f_{XY}(x,y)\, dx\, dy$ (continuous case)

2.5.1 Addition Theorem of expectation:

The mathematical expectation of the sum of random variables is equal to the sum of their expectations i.e., if X and Y are random Variables then E(X + Y) = E(X) + E (Y)

2.5.2 Multiplication Theorem of expectation:

The mathematical expectation of the product of random variables is equal to the product of their expectations i.e., if X and Y are independent random variables then E(XY) = E(X) x E (Y)

2.5.3 Expectation of a linear combination of random variables

Let $X_1, X_2, \ldots . X_n$ be n random variables and $a_1, a_2, a_3, \ldots a_n$ be constants then

$$E\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i E(X_i)$$

2.5.4 Conditional expected values

If (X, Y) is a two dimensional discrete random variable with joint probability mass function P_{ij}, then the conditional expectations of g(X,Y) are defined as follows:

$$E\{g(X,Y) / Y = Yj\} = \sum_i g(x_i, y_j)\, P(X = x_i / Y = y_j)$$

$$= \sum_i g(x_i, y_j) \frac{p_{ij}}{p_{*j}}$$

$$E\{g(X,Y) / X = X_j\} = \sum_i g(x_i, y_j) \frac{p_{ij}}{p_{i*}}$$

If (X, Y) is a two dimensional continuous RV with joint pdf f(x,y), then

$$E\{g(X,Y) / Y\} = \int_{-\infty}^{\infty} g(x, y) \times f(x / y)\, dx$$

$$E\{g(X,Y) / X\} = \int_{-\infty}^{\infty} g(x, y) \times f(y / x)\, dy$$

Conditional means are defined as

NOTES

$$\mu_{Y/X} = E(Y/X) = \int_{-\infty}^{\infty} y\, f(y/x)\, dy$$

$$\mu_{X/Y} = E(Y/X) = \int_{-\infty}^{\infty} x\, f(x/y)\, dx$$

Conditional Variances are defined as

$$\sigma^2_{Y/X} = E\{(Y - \mu_{Y/X})^2\} = \int_{-\infty}^{\infty} (y - \mu_{Y/X})^2\, f(y/x)\, dy$$

$$\sigma^2_{X/Y} = E\{(X - \mu_{X/Y})^2\} = \int_{-\infty}^{\infty} (y - \mu_{Y/X})^2\, f(x/y)\, dx$$

Note:
If X and Y are independent RVs then E(Y/X) = E(Y) and E(X/Y) = E(X)

2. 6 COVARIANCE

Let (X, Y) be a bivariate random variable. Then covariance of (X,Y) is defined as

$$\begin{aligned} Cov(X,Y) &= E[(X - \bar{X})(Y - \bar{Y})] \\ &= E[XY - X\bar{Y} - \bar{X}Y + \bar{X}\,\bar{Y}\,] \\ &= E[XY] - \bar{Y}E[X] - \bar{X}\,E[Y] + \bar{X}\,\bar{Y} \\ &= E[XY] - \bar{X}\,\bar{Y} - \bar{X}\,\bar{Y} + X\,Y \\ &= E[XY] - E[X]\,E[Y] \end{aligned}$$

$$Cov(X,Y) = E[XY] - E[X]\,E[Y]$$

Note:
If X and Y are independent then E[XY] = E[X] E[Y] and hence in this case Cov(X, Y) = 0. But the converse is not true. i.e., if Cov(X, Y) = 0 then X and Y need not be independent.

(i) Cov (aX, bY) = abCov (X,Y)

$$\begin{aligned} Cov(aX, bY) &= E[(aX)(bY)] - E(aX)\,E(aY) \\ &= ab\,E(XY) - ab\,E(X)\,E(Y) \\ &= ab[E(XY) - E(X)\,E(Y)] \\ &= ab\,cov(X,Y) \end{aligned}$$

(ii) Cov (X+a, Y+ b) = Cov (X,Y)

$$Cov(X+a, Y+b) = E[(X+a)(Y+b) - E(X+a)\,E(Y+b)]$$

NOTES

$= E[XY+bX+aY+ab] - [E(X)+a]\,[E(Y)+b]$
$= E(XY)+bE(X)+aE(Y)+ab-E(X)E(Y)-aE(Y)-bE(X)-ab$
$= E(XY)- E(X)E(Y) = Cov\,(X,Y)$

(iii) $Cov\,(aX+b, cY+b) = ac\,Cov\,(X,Y)$

$Cov\,(aX+b, cY+b) = E[(aX+b)(cY+d)] - E(aX+b)\,E(cY+d)]$
$= E[acXY+adX+bcY+bd] - [aE(X)+b]\,[cE(Y)+d]$
$= acE(XY)+ adE(X)+ bcE(Y)+ bd - acE(X)E(Y) - adE(X) - bcE(Y) - bd$
$= ac[E(XY)- E(X)E(Y)] = ac\,Cov\,(X,Y)$

(iv) $V(X_1+X_2) = V(X_1) + V(X_2) + 2Cov(X_1,X_2)$

$V(X_1+X_2) = E[(X_1+X_2)^2] - E[(X_1+X_2)^2]$
$= E(X_1^2+2X_1X_2+X_2^2) - [E(X_1)+E(X_2)]^2$
$= E(X_1^2+2E(X_1X_2)+E(X_2^2) - [E(X_1)]^2 - [E(X_2)]^2 - 2E(X_1)E(X_2)$
$= E(X_1^2) - [E(X_1)]^2 + E(X_2^2) - [E(X_2)]^2 + 2[E(X_1X_2) - E(X_1)\,E(X_{2q})]$
$= V(X_1) + V(X_2) + 2Cov\,(X_1X_2)$

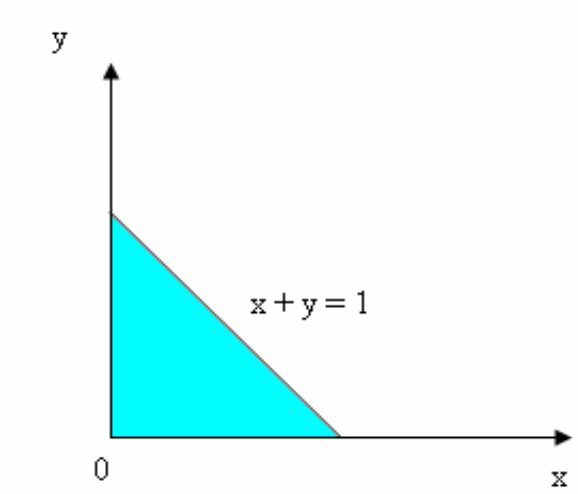
(v) $V(X_1-X_2) = V(X_1) + V(X_2) - 2Cov(X_1,X_2)$

If X_1 and X_2 are independent then

$V(X_1 \pm X_2) = V(X_1) + V(X_2)$

Example 1: The joint pdf of (X, Y) is given by f(x, y) = 24xy, x > 0, y > 0, x + y ≤ 1, and f(x,y) = 0 elsewhere, find the conditional mean and variance of Y given X.

Solution:



NOTES

$$f_X(x) = \int_0^{1-x} 24xy\,dxdy = 12x(1-x)^2,\ 0 < x < 1$$

$$f(y/x) = \frac{f(x,y)}{f_X(x)} = \frac{2y}{(1-x)^2},\ 0 < y < 1-x$$

$$E(Y/X = x) = \int_0^{1-x} y\,f(y/x)\,dy = \int_0^{1-x} \frac{2y^2}{(1-x)^2}\,dy = 2/3\,(1-x)$$

$$E(Y^2/x) = \int_0^{1-x} y^2 f(y/x)\,dy = \tfrac{1}{2}(1-x)^2$$

$$\text{Var}(Y^2/x) = E(Y^2/x) - [E(Y/x)]^2$$

$$= \tfrac{1}{2}(1-x)^2 - 4/9(1-x)^2 = 1/18(1-x)^2$$

Example 2: The jont distribution of (X,Y) is given by

X/ Y	1	3	9
2	1/8	1/24	1/12
4	1/4	1/4	0
6	1/12	1/24	1/12

find the Cov(X, Y)

Solution:

By definition cov(X,Y) = E[XY] - E[X] E[Y]

$$E(X) = \sum_i x_i\,p(x_i);\quad E(Y) = \sum_j y_j\,p(y_j)$$

Marginal probability distribution of X

X = i		$p_{i*} = \sum_{j=1,3,9} P_{ij}$
2	$P_{21} + P_{23} + P_{29}$	= 1/8 + 1/24 + 1/12 = 1/4
4	$P_{41} + P_{43} + P_{49}$	= ¼ + ¼ + 0 = ½
6	$P_{61} + P_{63} + P_{69}$	= 1/8 + 1/24 + 1/12 = 1/4

Marginal probability distribution of Y

NOTES

Y = j	$p_{*j} = \sum_{i=2,4,6} p_{ij}$	
1	$P_{21} + P_{41} + P_{61}$	=1/8 +1/4 + 1/8 = ½
3	$P_{23} + P_{43} + P_{63}$	= 1/24 + ¼ + 1/24 = **1/3**
9	$P_{29} + P_{49} + P_{69}$	= 1/12 +0 + 1/12 = **1/6**

Now $E(X) = \sum x_i p_{i*} = 2(1/4) + 4(1/2) + 6(1/4) = ¼$

$E(Y) = \sum y_j p_{*j} = 1(1/2) + 3(1/3) + 9(1/6) = 3$

$E(XY) = \sum\sum (x_i y_j) p(x_i, y_j) = 2.1\ (1/8) + 2.3\ (1/24) + 2.9\ (1/12) + 4.1\ (1/4)$
$+ 4.3\ (1/4) + 0 + 6.1\ (1/8) + 6.3\ (1/24) + 6.9\ (1/12)$
$= 9$

$Cov(X, Y) = E[XY] - E[X]\ E[Y]$
$= 9 - (3.4) = -3$

Example 3: Two random variables X and Y have joint pdf

$f_{XY}(x,y) = \frac{xy}{96}$, $0 < x < 4$, $1 < y < 5$
0, elsewhere

Find i) E(X) ii) E(Y) iii) E(XY) iv) E(2X + 3Y) v) Var(X) vi) Var(Y) and vii) Cov(X, Y)

Solution:

$$E(X) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x\, f(x,y)\,dx\,dy ; = \int_{1}^{5}\int_{0}^{4} x\,\frac{xy}{96}\,dx\,dy = 8/3$$

$$E(Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y\, f(x,y)\,dx\,dy = \int_{1}^{5}\int_{0}^{4} y\,\frac{xy}{96}\,dx\,dy = 31/9$$

$$E(XY) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (xy)\, f_{XY}(x,y)\,dx\,dy$$

$$= \int_{1}^{5}\int_{0}^{4} (xy)\,\frac{xy}{96}\,dx\,dy = 248/27$$

$E(2X + 3Y) = 2E(X) + 3\,E(Y) = 2(8/3) + 3(31/9) = 47/3$

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

NOTES

$$E(X^2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^2 f(x,y)dxdy ; = \int_{1}^{5}\int_{0}^{4} x^2 \left(\frac{xy}{96}\right) dxdy = 8$$

$$Var(X) = E(X^2) - [E(X)]^2 = 8 - (8/3)^2 = 8/9$$

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

$$E(Y^2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y^2 f(x,y)dxdy ; = \int_{1}^{5}\int_{0}^{4} y^2 \left(\frac{xy}{96}\right) dxdy = 13$$

$$Var(Y) = E(Y^2) - [E(Y)]^2 = 13 - (31/9)^2 = 92/81$$

$$Cov(X, Y) = E[XY] - E[X]\,E[Y]$$
$$= (248/27) - (8/3)(31/9) = 0$$

Since Cov(X, Y) = 0, X and Y are independent.

Try yourself !

1. The random variable (X, Y) has the following joint pdf
   $f(x, y) = \frac{1}{2}(x + y)$, $0 \le x \le 2$, $0 \le y \le 2$
   0, otherwise
   a) Obtain the marginal distributions of X.
   b) E(X) and $E(X^2)$
   c) Compute Covariance (X,Y)
   (Solution: f(x) = x + 1; f(y) = (1 + y); E(X) = 14/3, E(Y) = 14/3; $E(X^2)$ = 20/3, Cov(X,Y) = -60)

2. If X and Y are discrete random variables with P[X = x, Y = y] = C(x + y) for x = 0,1,2 and y = 1,2 and P[x, y] = 0, otherwise, find C and the covariance between X and Y.

(Solution: C = 1/15, Cov(X, Y) = -2 / 75)

2.7 CORRELATION

In probability theory and statistics, **correlation**, also called **correlation coefficient**, indicates the strength and direction of a linear relationship between two random variables. In general statistic usage, *correlation* or co-relation refers to the departure of two variables from independence.

To examine whether the two R.V.'s are inter-related, we collect n pairs of values of X and Y corresponding to n repetitions of the random experiment. Let them be (x_1, y_1), (x_2, y_2), …..

NOTES

(x_n, y_n). Then we plot the points with the co-ordinates (x_1, y_1), (x_2, y_2), …..(x_n, y_n) on a graph paper. The simple figure consisting of the plotted points is called a scatter diagram.From the scatter diagram, we can form a fairly good, though vague idea of the relationship between X and Y. If the points are dense or closely packed, we may conclude that X and Y are correlated. On the other hand if the points are widely scattered throughout the graph paper. We may conclude that X and Y are either not correlated or poorly correlated.

A number of different coefficients are used for different situations. The best known is the Pearson product-moment correlation coefficient, which is obtained by dividing the covariance of the two variables by the product of their standard deviations. Despite its name, it was first introduced by Francis Galton

2.7.1 Correlation coefficient : The correlation coefficient between two random variables X and Y is defined as

$$r = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E(XY) - E(X)E(Y)}{\text{vVar}(X)\ \text{v Var}(Y)}$$

2.7.2 Definition : Two random variables are uncorrelated with each other, if the correlation between X and Y is equal to the product of their means.

$E[XY] = E(X) . E(Y)$

2.7.3 Definition: The random variables are ***orthogonal*** to each other if the correlation between X and Y is equal to zero

Note:
Correlation coefficient is a number that varies between -1 and +1. When the correlation coefficient is 1 or -1, the random variables are perfectly correlated or have linear relationship between them. When r is 1, they are positively correlated. When r is -1, they are negatively correlated. If the correlation coefficient is 0, then the random variables are uncorrelated. If the two random variables are statistically independent then they are also uncorrelated and the covariance is zero but converse is not true.

2.7.4 KARL PEARSON FORMULAE

We have the following formulae for calculating the correlation coefficient between two variables x and y, with number of data n,

1. $$r_{XY} = \frac{n \Sigma xy - \Sigma x \Sigma y}{\text{v}\{n \Sigma x^2 - (\Sigma x)^2\}\text{v}\{n \Sigma y^2 - (\Sigma y)^2\}}.$$

NOTES

If no assumed average is taken for x and y series.

2. $r_{XY} = \frac{\Sigma dxdy}{\sqrt{\Sigma dx^2}\sqrt{\Sigma dy^2}}$ (Where dx = x – mean of x, dy = y – mean of y)

3. $r_{XY} = \frac{n \Sigma d\,xdy - \Sigma d\,x\,\Sigma\,dy}{\sqrt{\{n \Sigma d\,x^2 - (\Sigma d\,x)^2\}}\sqrt{\{n \Sigma\,dy^2 - (\Sigma\,dy)^2\}}}$.

This formula is used when deviations for x and y series are taken from some assumed values. dx = x – A and dy = y – B.

4. In a bivariate frequency distribution of variables x and y correlaton coefficient is given by

$r_{XY} = \frac{N \Sigma f\,dxdy - \Sigma fdx\,\Sigma fdy}{v\{N \Sigma fdx^2 - (\Sigma fdx)^2\}v\{N\Sigma fdy^2 - (\Sigma fdy)^2\}}$.

Note:
1.Correlation coefficient is independent of change of origin and scale.
ie., If $U = \frac{X - A}{h}$ and $V = \frac{Y - B}{k}$ where h, k>0, then $r_{XY} = r_{UV}$
If X and Y take considerably large values, computation of r_{XY} will become difficult. In such problems, we may introduce change of origin and scale and compute r using the above property.

2. $r_{XY} = \frac{\sigma x^2 + \sigma_Y{}^2 - \sigma^2_{(x-y)}}{2\,\sigma x\ \sigma_Y}$

2.7. 5 Spearman's rank correlation

In statistics, **rank correlation** is the study of relationships between different rankings on the same set of items. It deals with measuring correspondence between two rankings, and assessing the significance of this correspondence.

The Karl Pearson's formula for calculating r is developed on the assumption that the values of the variables are exactly measurable. In some situations, it may not be possible to give precise values for the variables. In such cases we rank the observations in ascending or descending order using the numbers 1,23, ... n and measure the degree of relationship between the ranks instead of actual numerical values. The rank correlation coefficient when there are n ranks in each variable is given by formula

$\rho = 1 - \frac{6 \Sigma d^2}{n(n^2-1)}$

NOTES

where d = x- y is the difference between ranks of corresponding pairs x and y.
n = number of observation.

2.7.5.1 Tie ranks: When the values of variables x and y are given, we can rank the values in each of the variables x and y are given, we can rank the values in each of the variables and determine the Spearman's rank correlation coefficient. If two or more observations have the same rank we assign to them the mean rank. In this case, there is a correlation factor in the formula for ρ. The formula for V is given by,

$$\rho = 1 - \frac{6[\ \Sigma d^2 + \Sigma m(m^2 - 1)/12\ ]}{n\,(n^2-1)}$$

where m denotes the number of times the rank is repeated .

Example 1: Two R.V X and Y take the values 1,2,3 and their probabilities are as follows:

Y \ X	1	2	3	f(y)
1	0.1	0.1	0.1	0.3
2	0.1	0.2	0.1	0.4
3	0.1	0.1	0.1	.3
f(x)	0.3	0.4	0.3	1

Find mean, variance , E(X + Y).and correlation of X and Y (r_{XY}).

Solution:

Y \ X	1	2	3	p_{*j} (= f(y))
1	0.1	0.1	0.1	0.3
2	0.1	0.2	0.1	0.4
3	0.1	0.1	0.1	.3
p_{i*} (= f(x))	0.3	0.4	0.3	1

Now E(X) = $\sum x_i p_{i*}$ = (1 x 0.3) + (2 x 0.4) + (3 x 0.3) = 2

Therefore mean of X = 2

Now E(Y) = $\sum y_j p_{*j}$ = (1 x 0.3) + (2 x 0.4) + (3 x 0.3) = 2

Therefore mean of Y = 2

NOTES

Now

$E(X^2) = \sum x_i^2 p_{i*} = (1^2 \times 0.3) + (2^2 \times 0.4) + (3^2 \times 0.3) = 4.6$

$E(Y^2) = \sum y_j^2 p_{*j} = (1^2 \times 0.3) + (2^2 \times 0.4) + (3^2 \times 0.3) = 4.6$

$Var(X) = E(X^2) - [E(X)]^2 = 4.6 - 4 = 0.6$

$Var(Y) = E(Y^2) - [E(Y)]^2 = 4.6 - 4 = 0.6$

$E(X + Y) = E(X) + E(Y) = 2 + 2 = 4$

Now $Cov(X, Y) = E[XY] - E[X]\,E[Y]$

$E(XY) = \sum\sum (x_i y_j) p(x_i, y_j) = 1 \times 1 \times 0.1 + 1 \times 2 \times 0.1 + 1 \times 3 \times 0.1 + 2 \times 1 \times 0.1 + 2 \times 2 \times 0.2 + 2 \times 3 \times 0.1 + 3 \times 1 \times 0.1 + 3 \times 2 \times 0.1 + 3 \times 3 \times 0.1 = 4$

$Cov(X, Y) = E[XY] - E[X]\,E[Y] = 4 - 4 = 0$

Correlation of X and Y

$$r_{XY} = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$

$Cov(X, Y) = E[XY] - E[X]\,E[Y]$

$\sigma_X = \sqrt{var(X)}$: $Var(X) = E(X^2) - [E(X)]^2$

$\sigma_Y = \sqrt{var(Y)}$: $Var(Y) = E(Y^2) - [E(Y)]^2$

Since $cov(X, Y) = 0$, $r_{XY} = 0$

Example 2: Suppose that 2 random variables (X, Y) has the joint pdf

$f(x.y) = x + y, \quad 0 < x < 1, 0 < y < 1$

$0, \quad$ otherwise

Obtain the correlation coefficient between X and Y ?

Solution :

Correlation of X and Y

$$r_{XY} = \frac{cov(X, Y)}{\sigma_X \sigma_Y}$$

$Cov(X, Y) = E[XY] - E[X]\,E[Y]$

$\sigma_X = vvar(X)$: $Var(X) = E(X^2) - [E(X)]^2$

NOTES

σ_Y = vvar(Y) : Var (Y) = $E(Y^2) - [E(Y)]^2$

$$E(X) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x\, f(x,y)dxdy\ ; = \int_{0}^{1}\int_{0}^{1} x\,(x+y)\,dxdy = 7/12$$

$$E(Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y\, f(x,y)dx\,dy = \int_{0}^{1}\int_{0}^{1} y\,(x+y)\,dxdy = 7/12$$

$$E(X^2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^2\, f(x,y)dxdy\ ; = \int_{0}^{1}\int_{0}^{1} x^2\,(x+y)\,dxdy = 5/12$$

$$E(Y^2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y^2\, f(x,y)dx\,dy = \int_{0}^{1}\int_{0}^{1} y^2\,(x+y)\,dxdy = 5/12$$

$$E(XY) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x\,y\, f(x,y)dx\,dy = \int_{0}^{1}\int_{0}^{1} xy\,(x+y)\,dxdy = 1/3$$

Var (X) = $E(X^2) - [E(X)]^2$ = 5/12 – 49/144 = 11/144

Var (Y) = $E(Y^2) - [E(Y)]^2$ = 5/12 – 49/144 = 11/144

$\sigma_X = \sqrt{var(X)} = \sqrt{(11/144)}$

$\sigma_Y = \sqrt{var(Y)} = \sqrt{(11/144)}$

$$r_{XY} = \frac{Cov(X, Y)}{\sigma_X \sigma_Y} = \frac{E[XY] - E[X]\,E[Y]}{\sigma_X \sigma_Y}$$

Cov(X, Y) = (1/3) – (7/12)(7/12) = -1/144

$$r_{XY} = \frac{cov(X, Y)}{\sigma_X \sigma_Y} = \frac{-1/144}{\sqrt{(11/144)}\,\sqrt{(11/144)}}$$

Example 3: The random variable X has a mean value 3 and variance 2. A new random variable Y is defined as Y = 3X – 11. Check whether a) X and Y are orthogonal to each other b) X and Y are uncorrelated to each other.

Solution :

Given E(X) = 3 and σ^2_X = 2

Mean value of the random variable Y = E(3X – 11) = 3E(X) – 11 = 3 x 3 – 11= -2

NOTES

Correlation of X and Y is E(XY) $= E(X(3X - 11)) = E(3X^2 - 11X)$
$= 3E(X^2) - 11\,E(X)$

(Add and subtract $3[E(X)]^2$)

$= 3E(X^2) - 11\,E(X) + 3[E(X)]^2) - 3[E(X)]^2)$
$= \{3E(X^2) - 3[E(X)]^2\} + 3[E(X)]^2 - 11\,E(X)$
$= 3\,Var\,(X) + 3[E(X)]^2 - 11\,E(X)$
$= 3 \times 2 + 3(3^2) - 11 \times 3 = 6 + 27 - 33$
$= 0$

Therefore $E(XY) = 0$
Since $E(XY) = 0$, X and Y are orthogonal.

But $E(XY) \neq E[X]\,E[Y]$, they are correlated.

Example 4: Two random variables X and Y are defined as $Y = 4X + 9$. Find the correlation coefficient between X and Y.
Solution:

$$r_{XY} = \frac{cov(X, Y)}{\sigma_X \sigma_Y} = \frac{E[XY] - E[X]\,E[Y]}{\sigma_X \sigma_Y}$$

$$= \frac{E[X(4X + 9)] - E[X]\,E[4X + 9]}{\sigma_X \sigma_Y} = \frac{4\,E(X^2) + 9\,E(X) - 4[E(X)]^2 - 9\,E(X)}{\sigma_X \sigma_Y}$$

$$= \frac{4\{\,E(X^2) - [E(X)]^2\,\}}{\sigma_X \sigma_Y} = \frac{4\,\sigma_X^{\,2}}{\sigma_X \sigma_Y} = \frac{4\,\sigma_X}{\sigma_Y}$$

Let us find the standard deviation of Y, namely σ_Y

$\sigma_Y^{\,2} = E(Y^2) - [E(Y)]^2$
$= E(4X + 9)^2) - [E(4X + 9)]^2 = E(16\,X^2 + 72\,X + 81) - [4\,E(X) + 9]^2$
$= 16\,E(X^2) + 72\,E(X) + 81 - 16[E(X)]^2 - 72\,E(X) - 81$
$= 16\,E(X^2) - 16[E(X)]^2$
$= 16\,\sigma_X^{\,2}$

Therefore $r = \frac{4\,\sigma_X}{\sigma_Y} = \frac{4\,\sigma_X}{4\,\sigma_X} = 1$

Correlation coefficient r = 1
As the correlation coefficient is 1, they are positively perfectly correlated.

Example 5 : Let the random variable X have the marginal density $f(x) = 1, -\frac{1}{2} < x < \frac{1}{2}$ and let the conditional density of Y given X, be
$f(y / x) = 1, x < y < x + 1, -\frac{1}{2} < x < 0$

NOTES

$= 1, x < y < x + 1, -½ < x < 0.$

Show that the variables X and Y are uncorrelated.

Solution:

We have $E(X) = \int_{-½}^{½} f(x)\,dx = \int_{-½}^{½} x\,dx = 0$

If f(x, y) is the joint pdf of X and Y, then

$f(x, y) = f_X(x) \cdot f(y/x)$

$$E(XY) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, f(x,y)\,dx\,dy$$

$$= \int_{-½}^{0}\int_{x}^{x+1} xy\,dx\,dy + \int_{0}^{½}\int_{-x}^{1-x} xy\,dx\,dy$$

$$= ½ \int_{-½}^{0} x\left[y^2\right]_{x}^{x+1} dx + ½ \int_{0}^{½} x\left[y^2\right]_{-x}^{1-x} dx$$

$= 0$

Therefore Cov(X, Y) = E(XY) – E(X)E(Y)

$= 0$

Therefore the coefficient of correlation r = 0

$\Rightarrow$ the variables X and Y are uncorrelated.

Example 6: If X, Y and Z ate uncorrelated RVs with mean zero and SD 5,12 and 9 respectively, and if U = X + Y and V = Y+ Z, find the correlation coefficient between U and V.

Solution:

Since X, Y and Z are uncorrelated,

Cov(X, Y) = 0 ;Cov(Y,Z) = 0; Cov(Z,X) = 0

i.e., E(XY) – E(X)E(Y) = 0,

E(YZ) – E(Y)E(Z) = 0,

E(ZX) – E(Z)E(X) = 0

Which means

E(XY) = E(X)E(Y)

E(YZ) = E(Y)E(Z)

E(ZX) = E(Z)E(X)

Given E(X) = 0, E(Y) = 0, E(Z) = 0

Therefore from the above equation we will have

E(XY) = 0; E(YZ) = 0; E(ZX) = 0

NOTES

Given SD, $\sigma_X = 5$; $\sigma_Y = 12$; $\sigma_Z = 9$

⇨ Var(X) = 25; Var(Y) = 144; Var(Z) = 81

⇨ $E(X^2) = 25$; $E(Y^2) = 144$; $E(Z^2) = 81$ $(Var(X)) = E(X^2) - [E(X)]^2$ but $E(X) = 0$ hence $Var(X) = E(X^2)$)

The correlation coefficient $r_{UV} = \frac{Cov(U, V)}{\sigma_U \sigma_V}$

$Cov(U, V) = E(UV) - E(U)E(V)$

$E(UV) = E((X + Y)(Y + Z)) = E(XY + XZ + Y^2 + YZ)$
$= E(X)E(Y) + E(X)E(Z) + E(Y^2) + E(Y).E(Z)$

$E(U)E(V) = E(X + Y).E(Y + Z)$
$= [E(X) + E(Y)][E(Y) + E(Z)]$
$= E(X)E(Y) + E(X)E(Z) + [E(Y)]^2 + E(Y).E(Z)$

$Cov(U, V) = E(Y^2) - [E(Y)]^2 = Vary = 144$

$\sigma_U^2 = Var(U) = Var(X + Y) = Var(X) + Var(Y) + 2Cov(X,Y)$
$= 25 + 144 + 0 = 169$

⇨ $\sigma_U = 13$

$\sigma_V^2 = Var(V) = Var(Y + Z) = Var(Y) + Var(Z) + 2Cov(Y,Z)$
$= 144 + 81 + 0 = 225$

⇨ $\sigma_V = 15$

Therefore $r_{UV} = \frac{144}{13 \times 15} = 0.7385$

Example7: let X_1 and X_2 be two independent random variables with mean 5 & 10 and SDs 2 and 3 respectively. Obtain r_{UV} where $U = 3X_1 + 4X_2$ and $V = 3X_1 - X_2$

Solution:

$E(X_1) = 5$; $E(X_2) = 10$; SD = 2 & 3

$V(X_1) = 4$ and $V(X_2) = 9$

Given $U = 3X_1 + 4X_2$ and $V = 3X_1 - X_2$

$E(U) = E(3X_1 + 4X_2) = 3E(X_1) + 4E(X_2) = 3 \times 5 + 4 \times 10 = 55$

$E(V) = E(3X_1 - X_2) = 3E(X_1) - E(X_2) = 3 \times 5 - 10 = 5$

NOTES

$E(UV) = E[(3X_1 + 4X_2)(3X_1 - X_2)] = E(9X_1^2 + 12X_1X_2 - 3X_1X_2 - 4X_2^2)$
$= 9E(X_1^2) + 9E(X_1X_2) - 4E(X_2^2)$

$E(X_1X_2) = E(X_1)E(X_2)$ (X_1 and X_2 are independent)
$= 5 \times 10 = 50$

We have $V(X_1) = 4$

$E(X_1^2) - [E(X_1)]^2 = 4$
⇨ $E(X_1^2) = 4 + [E(X_1)]^2 = 4 + 25 = 29$
⇨ $E(X_2^2) = 9 + [E(X_2)]^2 = 9 + 100 = 109$
Therefore $E(UV) = 9 \times 29 + 9 \times 50 - 4 \times 109 = 275$

$$r_{UV} = \frac{Cov(U, V)}{\sigma_U \sigma_V}$$

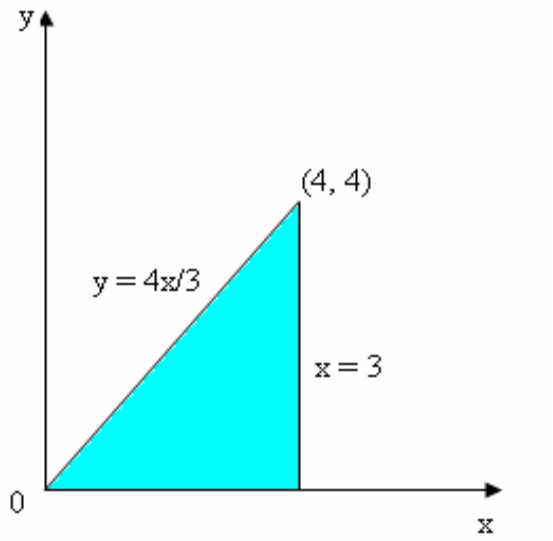
$Cov(U, V) = E(UV) - E(U)E(V) = 275 - 55 \times 5 = 0$

Therefore $r_{UV} = 0$

Example 8: If (X, Y) is two dimensional RV uniformly distributed over the triangular region R bounded by $y = 0$, $x = 3$ and $y = 4/3\ x$. Find $f_X(x)$, $f_Y(y)$, E(X), Var(X), E(Y), Var(Y) and r_{XY}.

Solution:

Since (X, Y) is uniformly distributed, f(x, y) = a constant = k
Now $\int\int f(x, y)\, dx\, dy = 1$



NOTES

$$\Rightarrow \int_0^4 \int_{3y/4}^3 k\, dx\, dy = 1$$

$$\Rightarrow 6k = 1$$

$$\Rightarrow k = 1/6$$

$$f_Y(y) = \int_{3y/4}^3 1/6\, dx = \frac{(4 - y)}{8},\ 0 < y < 4$$

$$f_X(x) = \int_{3}^{4x/3} 1/6\, dy = \frac{2x}{9},\ 0 < x < 3$$

$$E(X) = \int x\, f_X(x)\, dx = \int_0^3 \frac{2x^2}{9}\, dx = 2$$

$$E(Y) = \int y\, f_Y(y)\, dy = \int_0^4 \frac{y(4 - y)}{8}\, dy = 4/3$$

$$E(X^2) = \int x^2 f_X(x)\, dx = \int_0^3 \frac{2x^3}{9}\, dx = 9/2$$

$$E(Y^2) = \int y\, f_Y(y)\, dy = \int_1^4 \frac{y^2 (4 - y)}{8}\, dy = 8/3$$

$$Var(X) = E(X^2) - [E(X)]^2 = 9/2 - 4 = 1/2$$

$$Var(Y) = E(Y^2) - [E(Y)]^2 = 8/3 - 16/9 = 8/9$$

$$E(XY) = \int_0^4 \int_{3y/4}^3 1/6\, dx\, dy = 3$$

$$\mathbf{r_{XY}} = \frac{E(XY) - E(X)E(Y)}{\sigma x\ \sigma_Y} = \frac{3 - 2 \times 4/3}{\sqrt{1/2} \times 8/9} = \frac{1}{2}$$

Example 9: f X and Y are uncorrelated random variables with variances 16 and 9 , find the correlation coefficient between X + Y and X- Y.

Solution:

Let U = X + Y and V = X – Y

We have to find $r_{UV} = \frac{Cov(U, V)}{\sigma_U \sigma_V}$

E(U) = E(X) + E(Y) and E(V) = E(X) – E(Y)

Cov(U, V) = E(UV) – E(U)E(V)

NOTES

$E(UV) = E((X + Y)(X - Y)) = E(X^2 - Y^2) = E(X^2) - E(Y^2)$

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

$Cov(U, V) = E(X^2) - E(Y^2) - [E(X)]^2 + [E(Y)]^2 = [E(X^2) - [E(X)]^2] - [E(Y^2) - [E(Y)]^2]$
$= Var(X) - Var(Y) = \sigma_x^2 - \sigma_Y^2$

$Var(U) = Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$

Since X and Y are uncorrelated $Cov(X, Y) = 0$

$Var(U) = Var(X) + Var(Y)$

$\sigma_U^2 = \sigma_x^2 + \sigma_Y^2$

$Var(V) = Var(X - Y) = Var(X) + Var(Y) - 2Cov(X, Y)$

$Var(V) = Var(X) + Var(Y)$

$\sigma_V^2 = \sigma x^2 + \sigma_Y^2$

$$r_{UV} = \frac{Cov(U, V)}{\sigma_U \sigma_V} = \frac{\sigma_x^2 - \sigma_Y^2}{\sqrt{\sigma_x^2 + \sigma_Y^2}\sqrt{\sigma_x^2 + \sigma_Y^2}} = \frac{\sigma_x^2 - \sigma_Y^2}{\sigma_x^2 + \sigma_Y^2}$$

$$= \frac{16 - 9}{16 + 9} = 7/25 = 0.28$$

Example10:

Compute the coefficient correlation between X and Y, using the following data :

X	1	3	5	7	8	10
Y	8	12	15	17	18	20

Solution:

x	y	x^2	y^2	xy
1	8	1	64	8
3	12	9	144	36
5	15	26	225	75
7	17	49	289	119
8	18	64	324	144
10	20	100	400	200
34	90	248	1446	582

NOTES

$$r_{XY} = \frac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{\{n\Sigma x^2 - (\Sigma x)^2\}}\sqrt{\{n\Sigma y^2 - (\Sigma y)^2\}}}.$$

$$= \frac{6 \times 582 - 34 \times 90}{\sqrt{\{6 \times 248 - (34)^2\}}\sqrt{\{6 \times 1446 - (90)^2\}}}.$$

$= 0.9879$

Example11: Find the coefficient of correlation for the following data

X	35	40	60	79	83	95
Y	17	28	30	32	38	49

Solution:

X	Y	dx	dy	dx^2	dy^2	dxdy
35	17	-30	-15	900	225	450
40	28	-25	-4	625	16	10
60	30	-5	-2	25	4	10
79	32	14	0	196	0	0
83	38	18	6	324	36	108
95	49	30	17	900	289	510
392	194	2	2	2970	570	1178

$\bar{X} = 392 / 6 = 65.33$ $\bar{Y} = 194/ 6 = 32.33$

$dx = x - 65$ $dy = y-32$

$$r = \frac{n\Sigma dxdy - \Sigma dx \Sigma dy}{\sqrt{n\Sigma dx^2 - (\Sigma dx)^2}\ \sqrt{n\Sigma dy^2 - (\Sigma dy)^2}}.$$

$$= \frac{6 \times 1178 - 2 \times 2}{\sqrt{6 \times 2970 - 2^2}\ \sqrt{6 \times 570 - 2^2}}.$$

$= 0.9055$

Example 12 : Find the coefficient of correlation between output and cost of an automobile factory from the following data

NOTES

Output of cars (in thousands)	3.5	4.2	5.6	6.5	7.0	8.2	8.8	9.0	9.7	10.0
Cost of cars (thousand Rs.)	9.8	9.0	9.8	8.4	8.3	8.2	8.2	8.0	8.0	8.1

The correlation coefficient is unaffected by the change of origin and the scale Multiply outputs by 10 and then subtract 35. Multiply the cost (in thousands of Rs.) by 10 and subtract 80.

Solution:

x	y	dx	dy	$(dx)^2$	$(dy)^2$	dxdy
0	18	-37	14	1369	196	-518
7	10	-30	6	900	36	-180
21	8	-16	4	256	16	-64
30	4	-7	0	49	0	0
35	3	-2	-1	4	1	2
47	2	10	-2	100	4	-20
53	2	16	-2	256	4	-32
55	0	18	-4	324	16	-72
62	0	25	-4	625	16	-100
65	1	28	-3	784	9	-84
375	48	5	8	4667	298	-1068

$dx = x-37 \quad dy = y - 4$

$$r = \frac{n \Sigma dxdy - \Sigma dx \Sigma dy}{\sqrt{n \Sigma dx^2 - (\Sigma dx)^2}\ \sqrt{n \Sigma dy^2 - (\Sigma dy)^2}} .$$

$$= \frac{10 \times (-1068) - 5 \times 8}{\sqrt{10 \times 4667 - 25}\ \sqrt{10 \times 298 - 64}} .$$

$= - 0.92$

Example 13: The following table gives the bivariate frequency distribution of marks in an intelligence test obtained by 100 students according to their age :

NOTES

Age (x) in years Marks (y)	18	19	20	21	Total
10-20	4	2	2	0	8
20-30	5	4	6	4	19
30-40	6	8	10	11	35
40-50	4	4	6	8	22
50-60	0	2	4	4	10
60-70	0	2	3	1	6
Total	19	22	31	28	100

Calculate the coefficient of correlation between age and intelligence.

Solution : Since the frequencies of various values of x and y are not equal to 1 each the formula for the computation of r_{XY} is taken with a slight modification as given –

$$r_{XY} = \frac{N \Sigma f\,dxdy - \Sigma fdx\,\Sigma fdy}{\sqrt{\{N\Sigma fdx^2 - (\Sigma fdx)^2\}}\sqrt{\{N\Sigma fdy^2 - (\Sigma fdy)^2\}}}.$$

Where $dx = x-20 = u$, $dy = \frac{y-35}{10} = v$,

$$r_{XY} = r_{UV} \quad \frac{N \Sigma f\,uv - \Sigma fu\,\Sigma fv}{\sqrt{\{N\Sigma fu^2 - (\Sigma fu)^2\}}\sqrt{\{N\Sigma fv^2 - (\Sigma fv)^2\}}}.$$

Calculation of Coefficient of Correlation :

y ↓ x →			18	19	20	21	Total f	fv	fv^2	fuv
	Mid Points		18	19	20	21				
		v ↓ u →	-2	-1	0	1				
10-20	15	-2	4 (4)	2 (2)	2 (0)	0 (-2)	8	-16	32	20
20-30	25	-1	5 (2)	4 (1)	6 (0)	4 (-1)	19	-19	19	10
30-40	35	0	6 (0)	8 (0)	10 (0)	11 (0)	35	0	0	0
40-50	45	1	4 (-2)	4 (-1)	6 (0)	8 (1)	22	22	22	-4
50-60	55	2	0 (-4)	2 (-2)	4 (0)	4 (2)	10	20	40	4
60-70	65	3	0 (-6)	2 (-3)	3 (0)	1 (3)	6	18	54	-3
Total f			19	22	31	28	**N = 100**	**25**	**167**	**27**
fu			-38	-22	0	28	**-32**			
fu^2			76	22	0	28	**126**			
fuv			18	-6	0	15	**27**			

In the above table, the figures enclosed in the circles are the values of uv

NOTES

Σfuv for the first column of the table is computed as follows
= 4 x 4 + 5 x 2 + 6 x 0 + 4 x -2 + 0 x -4 + 0 x -6 = 18
Σfuv for the first row of the table is computed as follows
= 4 x 4 + 2 x 2 + 2 x 0 + 0 x -2 = 20

Similarly other Σfuv values are computed. Value of ΣΣfuv obtained as the total of the entries of the last column and as that of the last row must tally.

$$r_{XY} = r_{UV} = \frac{100 \times 27 - (-32) \times 5}{\sqrt{\{100 \times 126 - (-32)^2\}}\sqrt{\{100 \times 167 - (5)^2\}}}$$

$= 0.1897$

Example 14: The following are the ***ranks*** obtained by 10 students in Statistics and Mathematics

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

To what extent is the knowledge of students in the two subjects related?

Solution:

x	y	x-y = d	d^2
1	1	0	0
2	4	-2	4
3	2	1	1
4	5	-1	1
5	3	2	4
6	9	-3	9
7	7	0	0
8	10	-2	9
9	6	3	9
10	8	2	4
			36

The rank correlation is given by

$$p = 1 - \frac{6 \Sigma d^2}{n (n^2-1)}$$

NOTES

$1 - \frac{6 \text{ X } 36}{10 \text{ X } 99} = 1 - 0.219 = 0.781$

Example 15: Ten competitors in a beauty contest are ranked by three judges in the following order

First Judge	1	4	6	3	2	9	7	8	10	5
Second Judge	2	6	5	4	7	10	9	3	8	1
Third Judge	3	7	4	5	10	8	9	2	6	1

Use the method of rank correlation coefficient to determine which pair of judges have the nearest approach to common taste in beauty

Solution :

Let x,y,z denote the ranks by 1st, 2nd and 3rd judges respectively

z	y	z	dxy(x - y)	dyz(y-z)	dzx(z -x)	dxy^2	dyz^2	dzx^2
1	2	3	-1	-1	-2	1	1	4
4	6	7	-2	-1	-3	4	1	9
6	5	4	1	1	2	1	1	4
3	4	5	-1	-1	-2	1	1	4
2	7	10	-5	-3	-8	25	9	64
9	10	8	-1	2	1	1	4	1
7	9	9	-2	0	-2	4	0	4
8	3	2	5	2	6	25	1	36
10	8	6	2	2	4	4	4	16
5	2	1	4	0	4	16	0	16
						82	22	158

$\rho_{xy} = 1 - \frac{6 \Sigma dxy^2}{n(n^2-1)} = 1 - \frac{6 \text{ x } 82}{10 \text{ x } 99} = 0.503$

$\rho_{yz} = 1 - \frac{6 \Sigma dyz^2}{n(n^2-1)} = 1 - \frac{6 \text{ x } 22}{10 \text{ x } 99} = 0.867$

$\rho_{zy} = 1 - \frac{6 \Sigma dzx^2}{n(n^2-1)} = 1 - \frac{6 \text{ x } 158}{10 \text{ x } 99} = 0.04$

Since the rank correlation between y and z is positive highest among the three coefficients, judges y and z have the nearest approach for common taste in beauty.

NOTES

Example 16 : Find the rank correlation coefficient for the following data

x	92	89	87	86	86	77	71	63	53	50
y	86	83	91	77	68	85	52	82	37	57

Let R_1 and R_2 denote the ranks in x and y respectively

In this X series 86 is repeated twice which are in the positions 4^{th} and 5^{th} ranks. Thus common rank 4.5 (which is the average of 4 and 5) is to be given for each 86.

Solution:

x	y	R_1	R_2	$d = R_1-R_2$	d^2
92	86	1	2	1	1
89	83	2	4	2	4
87	91	3	1	2	4
86	77	4.5	6	1.5	2.25
86	68	4.5	7	2.5	6.25
77	85	6	3	3	9
71	52	7	9	2	4
63	82	8	5	3	9
53	37	9	10	1	1
50	57	10	8	2	4
					44.50

$$\rho = 1 - \frac{6[\Sigma d^2 + \Sigma m(m^2-1)/12]}{n(n^2-1)}$$

$$= 1 - \frac{6[44.5 + 2(2^2-1)/12]}{10 \times 99} = 1 - \frac{6 \times 45}{990} = 0.727$$

$\rho = 0.727$

Example 17: Calculate the correlation coefficient for the following ages of the husbands(X) and wives(Y), using only standard deviations of X and Y:

X	23	27	28	28	29	30	31	33	35	36
Y	18	20	22	27	21	29	27	29	28	29

Solution:

NOTES

x	y	dx = x-30	dy = y-24	dx^2	dy^2	d = x-y	d^2
23	18	7	6	49	36	5	25
27	20	3	4	9	16	7	49
28	22	2	2	4	4	6	36
28	27	2	3	4	9	1	1
29	21	1	3	1	9	8	64
30	29	0	5	0	25	1	1
31	27	1	3	1	9	4	16
33	29	3	5	9	25	4	16
35	28	5	4	25	16	7	49
36	29	6	5	36	25	7	49
Total		0	10	138	174	50	306

$$\sigma_x^2 = \frac{\Sigma u^2}{n} - \left(\frac{\Sigma u}{n}\right)^2 = 138/10 = 13.8$$

$$\sigma_Y^2 = \frac{\Sigma v^2}{n} - \left(\frac{\Sigma v}{n}\right)^2 = 174/10 - (10/10)^2 = 16.4$$

$$\sigma(x - Y)^2 = \sigma_D^2 = \sigma_x^2 = \frac{\Sigma d^2}{n} - \left(\frac{\Sigma d}{n}\right)^2 = 306/10 - (50/10)^2 = 5.6$$

$$r_{XY} = \frac{\sigma_x^2 + \sigma_Y^2 - \sigma^2_{(x-y)}}{2\,\sigma_x\,\sigma_Y} = \frac{13.8 + 16.4 - 5.6}{2\ v13.8 \times 16.4} = 0.8176$$

How you understood ?

Say true or false.

1. The variance of the sum of the random variables equals the sum of the variances if the random variables are uncorrelated.
2. $Cov(X_i, X_j)$ ¢ ¢ $Cov(X_j, X_i)$
3. If X_1 and X_2 are uncorrelated then X_1 and X_2 are not necessarily statistically independent.
4. If X_1 and X_2 are statistically independent then $Cov(X_1, X_2) = 0$

(Answers: 1.true,.2.false,3.true,4.true)

NOTES

Short answer questions

1.Define the correlation between two random variables.
2.When are two random variables orthogonal to each other?
3. Define covariance between two random variables.
4. Define correlation coefficient.

Try yourself!

1) The joint probability mass function of X and Y is given below: –

y \ x	-1	+1
0	1/8	3/8
1	2/8	2/8

Find correlation coefficient of (X,Y)

(Solution : $r_{XY} = \frac{-1}{3}$.)

2) Two random variables X and Y have the joint density

$f(x,y) = \begin{cases} 2 - x - y, & 0 < x < 1, 0 < y < 1 \\ 0 & \text{otherwise} \end{cases}$. Show that cor (X,Y) = $\frac{-1}{11}$.

3) Find the coefficient of correlation between industrial production and export using the following data –

Production (X)	55	56	58	59	60	60	62
Export (Y)	35	38	37	39	44	43	44

(Solution : r (X, Y) = 0.8981)

2.8 REGRESSION

A regression equation allows us to express the relationship between two (or more) variables algebraically. It indicates the nature of the relationship between two (or more) variables. In particular, it indicates the extent to which you can predict some variables by knowing others, or the extent to which some are associated with others.

Correlation is the study of the degree of relationship between two variables if the relationship exists. Regression is the study of the relationship between the variables. If Y is the dependant variable and X is independent variable the linear relationship suggested

NOTES

between the variables is called the regression equation of Y on X. This regression equation is used to estimate the value of Y corresponding to a known value of X. On the other hand, If X is the dependant variable and Y is the independent variable the linear relation expressing X in terms of Y is called the regression equation of X on Y. This is used to estimate the value of X corresponding to a known value of Y.

2.8.1 Regression lines

When the random variables X and Y are linearly correlated, the points plotted on the scatter diagram, corresponding to n pairs of observed values of X and Y, will have a tendency to cluster round a straight line. The straight line is called regression line.
If we treat X as the independent variable and hence assume that the values of Y depend on those of X, the regression line is called the regression line of Y on X.

Thus the equation of regression line of Y on X is

$$(Y - \overline{Y}) = r \frac{\sigma_Y}{\sigma_X} (X - \overline{X})$$

Similarly, if we assume that the values of X depend on those of the independent variable Y, the regression line of X on Y is obtained.

Thus the regression equation of X on Y is

$$(X - \overline{X}) = r \frac{\sigma_X}{\sigma_Y} (Y - \overline{Y})$$

The slope $r \frac{\sigma_Y}{\sigma_X}$ is called the regression coefficient of Y on X and is denoted by $b_{YX.}$

The slope $r \frac{\sigma_X}{\sigma_Y}$ is called the regression coefficient of X on Y and is denoted by $b_{XY.}$

Therefore the above regression lines can be written as

Regression line of Y on X as

$$(y - \overline{y}) = b_{YX} (x - \overline{x})$$

The regression equation of X on Y is

$$(x - \overline{x}) = b_{XY} (y - \overline{y})$$

NOTES

Note:

$$b_{YX} = r\frac{\sigma_Y}{\sigma_X} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}\frac{\sigma_Y}{\sigma_X} = \frac{Cov(X,Y)}{\sigma_X^2}$$

$$= \frac{n\Sigma xy - \Sigma x\Sigma y}{n\Sigma x^2 - (\Sigma x)^2}$$

When deviations are taken from some assumed values, we have

$$b_{YX} = \frac{n\Sigma dxdy - \Sigma dx\Sigma dy}{n\Sigma dx^2 - (\Sigma dx)^2}$$

Similarly

$$b_{XY} = \frac{n\Sigma xy - \Sigma x\Sigma y}{n\Sigma x^2 - (\Sigma x)^2}$$

When deviations are taken from some assumed values, we have

$$b_{XY} = \frac{n\Sigma dxdy - \Sigma dx\Sigma dy}{n\Sigma dy^2 - (\Sigma dy)^2}$$

Note:

1. The correlation coefficient is the geometric mean between the two regression coefficients

$$b_{YX} = r\frac{\sigma_Y}{\sigma_X}$$

$$b_{XY} = r\frac{\sigma_X}{\sigma_Y}$$

$$b_{YX} \times b_{XY} = r^2$$

$$r = \pm\sqrt{(b_{YX} \times b_{XY})}$$

r is positive if the two regression coefficients are positive. r is negative if the two regression coefficients are negative.

2.If one regression coefficient is greater than one, the other regression coefficient is less than one

NOTES

$r^2 = 1$

$b_{YX} \times b_{XY} = 1$

$b_{YX} = \frac{1}{b_{XY}}$

Therefore if one regression coefficient is greater than one, the other regression coefficient is less than one.

3.When there is perfect linear correlation between X and Y, viz., when $r_{xy} = \pm 1$, the two regression lines coincide.

Regression curves for the means

Regression curve of Y on X is $y = E(Y/X = x)$

Regression curve of X on Y is $y = E(X/Y = y)$

Example 1 : From the following data find

(i) The two regression equations

(ii) The coefficient between the marks in Economics and Statistics

(iii) The most likely marks in statistics when marks in Economics are 30

Marks in Economics x	25	28	35	32	31	36	29	38	34	32
Marks in Statistics y	43	46	49	41	36	32	31	30	33	39

Solution:

X	Y	dx	dy	dx^2	dy2	dxdy
25	43	-6	7	36	49	-42
28	46	-3	10	9	100	-30
35	49	4	13	16	169	52
32	41	1	5	1	25	5
31	36	0	0	0	0	0
36	32	5	-4	25	16	-20
29	31	-2	-5	4	25	10
38	30	7	-6	49	36	-42
34	33	3	-3	9	9	-9
32	39	1	3	1	9	3
320	380	10	20	150	438	-73

$\bar{x} = \frac{\Sigma x}{n} = \frac{320}{10} = 32$ $\bar{y} = \frac{\Sigma y}{n} = \frac{380}{10} = 38.$

NOTES

$$b_{YX} = \frac{n\Sigma dxdy - \Sigma dx \Sigma dy}{n\Sigma dx^2 - (\Sigma d x)^2)} = \frac{-73 \times 10 - 10 \times 20}{10 \times 150 - 100} = -0.6643$$

$$b_{XY} = \frac{n\Sigma dxdy - \Sigma dx \Sigma dy}{n\Sigma dy^2 - (\Sigma d y)^2)} = \frac{-73 \times 10 - 10 \times 20}{10 \times 438 - 400} = -0.2337$$

$b_{YX} = -0.6643$

$b_{XY} = -0.2337$

Thus the equation of regression line of Y on X is

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

$(y - 38) = -0.6643 (x - 32)$

$y = -0.6643x + 59.2576$

Similarly, the regression equation of X on Y is

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

$(x - 32) = -0.237 (y - 38)$

$x = -0.2337y + 40.8806$

Coefficient of correlation

$r^2 = b_{XY} \times b_{YX} = -0.6643 \times -0.2337 = 0.1552$

$r = \pm 0.394$

Now we have to find the most likely marks in statistics (y) when marks in economics (x) are 30
$y = -0.6643x + 59.2576$
Put $x = 30$
$y = -0.6643 \times 30 + 59.2576 = 39$
when $x = 30$, $y = 39$

Example 2: The two lines of regression are
$8x - 10y + 66 = 0$ ……..(A)
$40x - 18y - 214 = 0$ ……..(B)

NOTES

The variance of x is 9. Find the

(i) mean values of x and y

(ii) correlation coefficient between x and y

Solution:

(i) since both the lines of regression passes through the mean values x and y, the point $(\bar{x}, \bar{y})$ must satisfy the two given regression lines.

$8\bar{x} - 10\bar{y} = -66$ …….(1)

$40x - 18y = 214$ …….(2)

(1) X 5 $\Rightarrow 40\bar{x} - 50\bar{y} = -330$ …….(3)

(2) X 1 $\Rightarrow \underline{40x - 18y = 214}$ …….(4)

(3) – (4) $-32\bar{y} = -544$

$\bar{y} = 17$

Sub in (1) we get

$8\bar{x} - 10(17) = -66$

$x = 13$

Hence the mean values are given by $\bar{x} = 13$ and $\bar{y} = 17$

(ii) Let us suppose the equation (A) is the equation of line of regression of x on y

$8x = 10y - 66$ i.e. $x = \frac{10}{8} y - 66$

i.e. $b_{xy} = \frac{10}{8}$

and (B) is the equation of the line of regression of y on x

$18y = 40x - 214$ i.e. $y = \frac{40}{18} x - \frac{214}{18}$

i.e,. $b_{yx} = \frac{40}{18}$

WKT $r^2 = b_{xy} \times b_{yx} = \frac{10}{8} \frac{40}{18}$

$r^2 = 2.77$

$r \not< 1$

Thus what we choose here is wrong.

NOTES

Hence choose equation of (A) is the the equation of line of regression of y on x and (B) equation of line of regression of x on y.

(A) $\Rightarrow$ $10y = 8x + 66$ i.e. $y = \frac{8}{10}x + \frac{66}{10}$

i.e $b_{yx} = \frac{8}{10}$

(B) $40x = 18y + 214$ i.e. $x = \frac{18}{40}y + \frac{214}{40}$

i.e. $b_{xy} = \frac{18}{40}$

WKT $r^2 = bxy \times byx$

$r^2 = \frac{8}{10}\,\frac{18}{40} = \frac{9}{25}$ $\quad r = \pm 0.6$

Since both the regression coefficients are positive r must be positive. Hence r = 0.6

Example 3: A study of prices of rice at Chennai and Madurai gave the following data:

	Chennai	Madurai
Mean	19.5	17.75
S.D.	1.75	2.5

Also the coefficient of correlation between the two is 0.8. Estimate the most likely price of rice (i) at Chennai corresponding to the price of 18 at Madurai and (ii) at Madurai corresponding the price of 17 at Chennai.

Solution:
Let the price of rice at Chennai and Madurai be denoted by X and Y respectively. Then from the data

$\overline{x} = 19.5, \overline{y} = 17.75, \sigma_X = 1.75, \sigma_Y = 2.5$ and $r_{XY} = 0.8$

Thus the equation of regression line of Y on X is

$(y - \overline{y}) = r\frac{\sigma_Y}{\sigma_X}(x - \overline{x})$

$y - 17.75 = \frac{0.8 \times 2.5}{1.75}(x - 19.5)$

NOTES

When x = 17,

$$y = 17.75 + \frac{0.8 \times 2.5}{1.75} \times (-2.5)$$
$$= 14.89$$

The regression equation of X on Y is

$$(x - \bar{x}) = \frac{\sigma_Y}{\sigma_X} (y - \bar{y})$$
$$x - 19.5 = \frac{0.8 \times 1.75}{2.5} (y - 17.75)$$

When y = 18

$$x = 19.5 + \frac{0.8 \times 1.75}{2.5} (18 - 17.75)$$
$$= 19.64$$

Example 4 : In a partially destroyed laboratory record of an analysis of correlation data. The data following results only are legible: Variance of X = 1. The regression equations are 3x + 2y – 26 and 6x + y = 31. What were (i) the mean values of X and Y? (ii) the standard deviation of Y? and (iii) the correlation coefficient between X and Y?

Solution:

(i) since the lines of regression intersect at $(\bar{x}, \bar{y})$, we have $3\bar{x} + 2\bar{y} = 26$ and $6\bar{x} + \bar{y} = 31$. Solving these equations, we get $\bar{x} = 4$ and $\bar{y} = 7$.

(ii) which of the two equations is the regression equation of Y on X and which one is the regression equation of X on Y are not known.

Let us tentatively assume that the first equation is the regression line of X on Y and second equation is the regression line of Y on X. Based on this assumption, the first equation can be re-written as

$$x = -\frac{2}{3} y + \frac{26}{3} \qquad (1)$$

and the other as $y = 6x + 31$ (2)

Then $b_{xy} = -\frac{2}{3}$ and $b_{yx} = -6$

Thus $r^2_{XY} = b_{xy} \times b_{yx} = 4$

NOTES

Thus $r_{XY} = -2$ which is wrong

Hence our tentative assumption is wrong.
Thus the first equation is the regression line of Y on X and re-written as

$$y = -\frac{3}{2}x + 13 \qquad (3)$$

The second equation is the regression line of X on Y and re-written as

$$x = -\frac{1}{6}y + \frac{31}{6} \qquad (4)$$

Hence the correct $b_{YX} = -\frac{3}{2}$ and the correct $b_{XY} = -\frac{1}{6}$

Thus $r^2_{XY} = b_{xy} \times b_{yx} = \frac{1}{4}$

Thus $r_{XY} = -\frac{1}{2}$ (both b_{xy} and b_{yx} are negative)

Now $\frac{\sigma^2_Y}{\sigma^2_X} = \frac{b_{yx}}{b_{xy}} = \frac{-3/2}{-1/6} = 9$

$\sigma^2_Y = 9 \times \sigma^2_X = 9$

$\sigma_Y = 3$

Example 5: Find the angle between two lines of regression. Deduce the condition for the two lines to be i) at right angles and ii) coincident.

Solution:
The equations of the regression lines are

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

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

Slope of the first line is $r\frac{\sigma_Y}{\sigma_X} = m_1$

Slope of the second line is $r\frac{\sigma_X}{\sigma_Y} = m_2$

If θ is the angle between the two lines , then

$$\tan\theta = \frac{|m_1 - m_2|}{1 + m_1 m_2}$$

NOTES

$$= \left| \frac{r\frac{\sigma_Y}{\sigma_X} - \frac{\sigma_Y}{r\sigma_X}}{1 + \frac{\sigma_Y^2}{\sigma_X^2}} \right|$$

$$= \frac{|r - 1/r|\,\sigma_X\,\sigma_Y}{\sigma_X^2 + \sigma_Y^2}$$

$$= \frac{(1 - r^2)}{|r|} \times \frac{\sigma_X\,\sigma_Y}{\sigma_X^2 + \sigma_Y^2}$$

The two regression lines are at right angles when $\theta = \pi/2$, i.e. $\tan\theta = \infty$
i.e, $r = 0$

Therefore when the linear correlation between X and Y is zero, the two lines of regression will be at right angles.

The two regression lines are coincident, when $\theta = 0$, i.e., when $\tan\theta = 0$
i.e., when $r = \pm 1$.

Therefore when the correlation between X and Y is perfect, t he two regression lines will coincide.

Example 6: In a bivariate population $\sigma_X = \sigma_Y = \sigma$ and the angle between the regression line is $\tan^{-1}3$, obtain the value of correlation coefficient.

Solution:
The angle θ between the regression lines is given by

$$\tan\theta = \frac{(1 - r^2)}{|r|} \times \frac{\sigma_X\,\sigma_Y}{\sigma_X^2 + \sigma_Y^2}$$

Using $\tan\theta = 3$ and $\sigma_X = \sigma_Y = \sigma$

$$\frac{(1 - r^2)}{r}\,\frac{\sigma^2}{2\sigma^2} = 3$$

$\Rightarrow r^2 + 6r - 1 = 0$

Solving, $r = \frac{-6 \pm 6.325}{2}$

Since $-1 < r <$, $r = 0.1625$

Example 7 : Calculate the co-efficient of correlation between x and y from the following table and write down the regression equation of y on x:

NOTES

y / x	0 -	40 -	80 -	120 -
10 –	9	4	1	0
30 -	47	19	6	0
50 -	26	18	11	0
70 -	2	3	2	2

Solution :

Before doing this problem you should remember certain things.

For **grouped data**

i.

ii. $r_{XY} = r_{UV}$

iii. $b_{XY} \overset{4}{4} b_{UV}$

$b_{XY} = b_{UV} \times (i_x/i_y)$, where i_x is the increment in x and i_y is the increment in x.

iii. $b_{YX} \overset{4}{4} b_{VU}$

$b_{YX} = b_{VU} \times (i_y/i_x)$, where i_x is the increment in x and i_y is the increment in x.

Mean $\overline{X} = A + \left[\frac{\Sigma fx}{N} \times c \right]$, where N is the total frequency, $N = \Sigma f$

A is the mid value and C is the increment.

Similarly $\overline{Y} = A + \left[\frac{\Sigma f_Y}{N} \times c \right]$

In this problem we will find the regression coefficients, then coefficient of correlation and then regression equation of y on x.

NOTES

Calculation of Coefficient of Correlation :

x → / y ↓			0-40	40-80	80-120	120-160	Total f	fv	fv^2	fuv
	Mid Points		20	60	100	140				
		u → / v ↓	-1	0	1	2				
10-30	20	-1	9 (1)	4 (0)	1 (-1)		14	-14	14	8
30-50	40	0	47 (0)	19 (0)	6 (0)		72	0	0	0
50-70	60	1	26 (-1)	18 (0)	11 (1)		55	55	55	-15
70-90	80	2	2 (-2)	3 (0)	2 (2)	2 (4)	9	18	36	8
Total f			84	44	20	2	**N =150**	**59**	**105**	**1**
fu			-84	0	20	4	**-60**			
fu^2			84	0	20	8	**112**			
fuv			-21	0	14	8	**1**			

In the above table, the figures enclosed in the circles are the values of uv

Σfuv for the first column of the table is computed as follows
= 9 x 1 + 47 x 0 + 26 x -1 + 2 x -2 = 84
Σfuv for the first row of the table is computed as follows
= 9 x 1 + 4 x 0 + 1 x -1 = 8

Similarly other Σfuv values are computed. Value of ΣΣfuv obtained as the total of the entries of the last column and as that of the last row must tally.

$$b_{XY} = b_{UV} \times (i_x/i_y)$$
$$= \frac{N\Sigma fuv - \Sigma fu\Sigma fv}{N\Sigma fv^2 - (\Sigma fv)^2} = \frac{150 \times 1 - -60 \times 59}{150 \times 105 - (59)^2} \times \left(\frac{40}{20}\right) = 0.6015$$

$$b_{YX} = b_{VU} \times (i_y/i_x)$$

$$= \frac{N\Sigma fuv - \Sigma fu\Sigma fv}{N\Sigma fu^2 - (\Sigma fu)^2} = \frac{150 \times 1 - -60 \times 59}{150 \times 112 - (-60)^2} \times \left(\frac{20}{40}\right) = 0.1398$$

Coefficient of correlation $r = \sqrt{b_{YX} \times b_{XY}} = \sqrt{0.6015 \times 0.1398} = 0.29$
Therefore r = 0.29

Now the regression equation of y on x.

$$(y - \bar{y}) = b_{YX} (x - \bar{x}) \qquad (1)$$

NOTES

$$\bar{x} = A + \left[\frac{\Sigma fu}{N} \times c \right]$$

$$= 60 + \left[\frac{-60}{150} \times 40 \right] = 44$$

$$\bar{y} = A + \left[\frac{\Sigma fv}{N} \times c \right]$$

$$= 40 + \left[\frac{59}{150} \times 20 \right] = 47.87$$

Regression equation of y on x.

$(y - 47.87) = 0.1398(x - 44)$

$Y = 0.1398 x + 41.7188$

How you understood ?

Say true or false.

1. Correlation between variables gives the degree of relationship between them.
2. Regression between the variables gives the degree of relationship between them.
3. Regression between X and Y is same as that between Y and X.

Short answer questions

1. Define regression of y on x.
2. What are regression lines.
3. Define regression coefficient.
4. Differentiate between regression and correlation.

Try yourself !

1) Given that $x = 4y + 5$ and $y = kx + 4$ are the regression lines of X on Y and Y on X respectively. Show that $0 \leq k \leq \frac{1}{4}$. If $k = \frac{1}{16}$, find the means of X and Y and r_{XY}.

(Solution : mean of x = 28, mean of y = 5.75, $r_{XY} = ½$)

2) If the equations of the two lines of regression of y on x and x on y are respectively, $7x - 16y + 9 = 0$; $5y - 4x - 3 = 0$. Calculate the coefficient of correlation, x and y .

(Solution: r = 0.7338, mean of x = -3/29 and mean of y = 15/29)

3) The following table gives the data on rainfall and discharge in a certain river. Obtain the line of regression of y on x.

NOTES

Rainfall (inches) (X):	1.53	1.78	2.60	2.95	3.42
Discharge (1000 cc) (Y):	33.5	36.3	40.0	45.8	53.5

(**Solution:** $y = 9.7992x + 17.714$)

4) The following table gives according to age X, the frequency of marks obtained Y by 100 students in an intelligence test. Measure the degree of relationship between age and intelligence test.

Age / Marks	16-17	17-18	18-19	19-20
30-40	20	10	2	2
40-50	4	28	6	4
50-60	0	5	11	0
60-70	0	0	2	0
70-80	0	0	0	5

(Solution: $r = 0.6137$)

2.9 TRANSFORMATION OF RANDOM VARIABLES

2.9.1 Two functions of two random variables

If (X, Y) is a two dimensional random varialble with joint pdf $f_{XY}(x, y)$ and if $Z = g(X, Y)$ and $W = h(X, Y)$ ae two other random variables then the joint pdf of (Z, W) is given by

$$f_{ZW}(z,w) = |J| f_{XY}(x, y)$$

$$J = \begin{vmatrix} \frac{\partial x}{\partial z} & \frac{\partial x}{\partial w} \\ \frac{\partial y}{\partial z} & \frac{\partial y}{\partial w} \end{vmatrix}$$

2.9.2 One function of two random variables

If a random variable Z is defined as $Z = g(X, Y)$, where X and Y are random variables with joint pdf f(x, y).To find the pdf of Z were introduce a second random variable

$W = h(x, y)$ and obtain the joint pdf of (Z, W) by using the previous result. Let it be $f_Z(z)(z, w)$. The required pdf of Z is then obtained as the marginal pdf which is obtained by simply integrating $f_{ZW}(z, w)$ wiyh respect to w.

NOTES

$$f_Z(z) = \int_{-\infty}^{\infty} f_{ZW}(Z, W)\, dw$$

Example1 : If X and Y are independent random variables with pdf e^{-x}, $x \geq 0$; e^{-y}, $y \geq 0$ respectively. Find the density function of $U = \frac{X}{X+Y}$, and $V = X + Y$. Are U and V independent?

Solution:
Since X and Y are independent,

$$f_{XY}(x, y) = e^{-x} e^{-y} = e^{-(x+y)},\ x, y \geq 0$$

Solving equations $u = \frac{x}{x+y}$ and $v = x + y$, we get,

$x = uv$ and $y = v(1 - u)$

$$J = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} v & u \\ -v & 1-u \end{vmatrix} = v(1 - u) + uv = v$$

The joint pdf of (U,V) is given by

$$f_{UV}(u,v) = |J| f_{XY}(x,y)$$
$$= v\, e^{-(x+y)}$$
$$= v\, e^{-(uv + v(1-u))}$$
$$= v\, e^{-v}$$

The range space of (U, V) is obtained as follows :
Since $x, y \geq 0$, $uv \geq 0$ and $v(1- u) \geq 0$

Therefore either $u \geq 0$, $v \geq 0$ and $1 - u \geq 0$
i.e., $0 \leq u \leq 1$ and $v \geq 0$,
or $u \leq 0$, $v \leq 0$, $1-u \leq 0$ i.e., $u \leq 0$, $u > 1$ which is absurd.

Therefore, the range space (U, V) is given by $0 \leq u \leq 1$ and $v \geq 0$

Therefore $f_{UV}(u,v) = v\, e^{-v}$, $0 \leq u \leq 1$ and $v \geq 0$

The pdf of U is given by ,

$$f_U(u) = \int_{-\infty}^{\infty} f_{UV}(u,v)\, dv$$

NOTES

$$= \int_{-\infty}^{\infty} v\, e^{-v}\, dv = 1$$ (using integration by parts $\int u\, dv = uv - \int v\, du$ here $u = v$ and $dv = e^{-v}$)

$f_U(u) = 1$

i.e., U is uniformly distributed in (0, 1)

The pdf of V is given by

$$f_V(v) = \int_{-\infty}^{\infty} f_{UV}(u,v)\, du$$

$$= \int_{-\infty}^{\infty} v\, e^{-v}\, dv = ve^{-v}, \; v = 0$$

Now $f_U(u)\, f_V(v) = v\, e^{-v} = f_{UV}(u,v)$

Therefore U and V are independent random variables.

Example 2: If X and Y are independent random variables with $f_X(x) = e^{-x}U(x)$ and $f_Y(y) = 3e^{-3y}U(y)$, find $f_Z(z)$, if $Z = X/Y$.

U(X) is the unit step impulse function defined as
$U(X) = 1$, if $x \overset{24}{12} = 0$
$= 0$, if $x < 0$

Solution:
Since X and Y are independent, $f_{XY}(xy) = 3e^{-(x+3y)}, \; x,y = 0$

Introduce the auxiliary variable $W = Y$ with $Z = X/Y$.
Therefore $x = zw$ and $y = w$

$$J = \begin{vmatrix} \frac{\partial x}{\partial z} & \frac{\partial x}{\partial w} \\ \frac{\partial y}{\partial z} & \frac{\partial y}{\partial w} \end{vmatrix}$$

$$J = \begin{vmatrix} w & z \\ 0 & 1 \end{vmatrix} = w$$

The joint pdf of (Z, W) is given by

NOTES

$f_{ZW}(z,w) = |J| f_{XY}(x, y)$
$= |w| \times 3e^{-(z+3)w}; z,w = 0$

The range space is obtained as follows:

Since $y \geq 0$ and $x \geq 0$ we have $w = 0$ and $zw = 0$

As $w = 0$, $z = 0$.

The pdf is the marginal pdf, obtained by integrating $f_{ZW}(z, w)$ with respect to w over the range of w

The pdf of z is given by

$$f_Z(z) = \int 3w\, e^{-(z+3)w}\, dw$$
$$= \frac{3}{(Z+3)^2} \cdot z \geq 0$$

Example 3: If X and Y follow an exponential distribution with parameter 1 and are independent, find the pdf of $U = X - Y$

Solution:

Since X and Y follow an exponential distribution with parameter 1

$f_X(x) = e^{-x}$, $x > 0$ and $f_Y(y) = e^{-y}$, $y > 0$

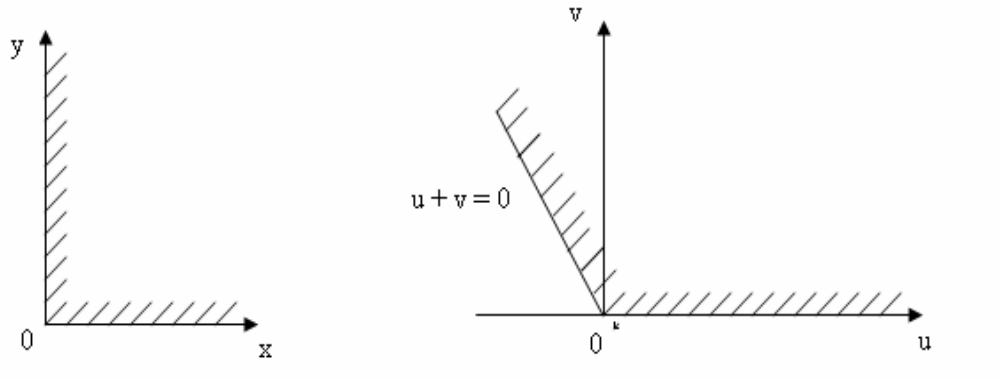
Since X and Y are independent $f_{XY}(x, y) = e^{-(x+y)}$; $x, y > 0$

Consider the auxiliary random variable $V = Y$ along with $U = X - Y$
$x = u + v$ and $y = v$

$$J = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} 1 & 1 \\ 0 & 1 \end{vmatrix} = 1$$

The joint pdf of (U, V) is given by
$f_{UV}(u,v) = |J| f_{XY}(x,y) = e^{-(x+y)} = e^{-(u+2v)}$
To find the range space of (U, V) under the transformation $x = u + v$ and $y = v$

NOTES



Therefore the range space of (U, V) is given by $v > -u$, when $u < 0$ and $v > 0$ when $u > 0$.

Now the pdf of U is given by

$$f_U(u) = \int_{-\infty}^{\infty} f_{UV}(u,v)\, dv$$

Therefore

$$f_U(u) = \int_{-u}^{\infty} e^{-(u+2v)}\, dv \quad \text{when } u < 0$$

$$= \int_{0}^{\infty} e^{-(u+2v)}\, dv \quad \text{when } u > 0$$

Therefore $f_U(u) = \frac{1}{2} e^{u}$, whaen $u < 0$

$= \frac{1}{2} e^{-u}$, when $u > 0$

Example 4: If X and Y are independent random variables each following N(0, 2), find the pdf of Z = 2X + 3Y

Solution:

Consider the auxiliary random variable W = Y along with Z = 2X + 3Y.

Therefore $z = 2x + 3y$ and $w = y$

$x = \frac{1}{2}(z - 3w)$ and $y = w$

$$J = \begin{vmatrix} \frac{\partial x}{\partial z} & \frac{\partial x}{\partial w} \\ \frac{\partial y}{\partial z} & \frac{\partial y}{\partial w} \end{vmatrix} = \begin{vmatrix} \frac{1}{2} & -3/2 \\ 0 & 1 \end{vmatrix} = \frac{1}{2}$$

Since X and Y are independent normal random variables

NOTES

$f_{XY}(x, y) = \frac{1}{8\pi} \exp[-(x^2 + y^2)/8]$, $-\infty < x, y < \infty$.

The joint pdf of (Z, W) is given by

$f_{ZW}(z, w) = |J| f_{XY}(x, y)$

$= ½ \times \frac{1}{8\pi} \exp[-\{(z - 3w)^2 + 4w^2\}/32]$, $-\infty < z, w < \infty$

The pdf of z is

$f_Z(z) = \frac{1}{16\pi} \int_{-\infty}^{\infty} \exp[-\{13 w^2 - 6zw + z^2/32]\ dw$

$= \frac{1}{(2\sqrt{13})\sqrt{2v\pi}} \exp\{-z^2 / 2(2\sqrt{13})^2\}$, $-\infty < z < \infty$, which is $N(0, 2\sqrt{13})$

Example 5: If X and Y are independent random variables each normally distributed with mean zero and variance σ^2, find the density functions of $R = v(X^2 + Y^2)$ and $\varphi = \tan^{-1}(Y/X)$.

Solution:
Since X and Y are independent $N(0, \sigma)$

$f_{XY}(x, y) = \frac{1}{2\pi \sigma^2} \exp[-(x^2 + y^2)/ 2 \sigma^2]$, $-\infty < x, y < \infty$.

$r = v(x^2 + y^2)$ and $\varphi = \tan^{-1}(y/x)$ are the transformations from Cartesian to polar coordinates.
Therefore , the inverse transformations are given by $x = r \cos\theta$ and $y = r \sin\theta$

$$J = \begin{vmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{vmatrix} = r$$

The joint pdf of (R, φ) is given by

$f_{R\varphi}(r, \theta) = \frac{|r|}{2\pi \sigma^2} \exp[- r^2/2 \sigma^2]$ $r \geq 0,\ 0 \leq \theta \leq 2\pi$

The density function of R is given by

$f_R(r) = \int_0^{2\pi} f_{R\varphi}(r, \theta)\, d\theta = \frac{r}{\sigma^2} \exp(- r^2/2 \sigma^2)$ $r \geq 0$

NOTES

The density function of φ is given by

$$f\varphi(\theta) = \int_0^{\infty} \frac{r}{2\pi\sigma^2} \exp[-r^2/2\sigma^2]\, dr$$

(put $t = r^2/2\sigma^2$)

$$= \frac{1}{2\pi} \int_0^{\infty} e^{-t}\, dt$$

$= 1/2\pi$, which is a uniform distribution.

Example 6: The joint pdf of a two dimensional random variable (X,Y) is given by

$f(x, y) = 4xy \exp[-(x^2 + y^2)]: x \geq 0, y \geq 0$
$= 0$ elsewhere
Find the density function $U = \sqrt{(X^2 + Y^2)}$

Solution:
Consider the auxiliary random variable V = Y along with $U = \sqrt{(X^2 + Y^2)}$

Therefore $x = \sqrt{(u^2 - v^2)}$ and $y = v \Rightarrow v = 0$, $u = 0$ and $u = v$
$\Rightarrow u \geq 0$ and $0 \leq v \leq u$

$$J = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} \frac{u}{\sqrt{(u^2 - v^2)}} & \frac{-v}{\sqrt{(u^2 - v^2)}} \\ 0 & 1 \end{vmatrix} = \frac{u}{\sqrt{(u^2 - v^2)}}$$

The joint pdf of (U, V) is given by

$$f_{UV}(u,v) = |J| f_{XY}(x,y) =$$

$$= \frac{u}{\sqrt{(u^2 - v^2)}} 4xy \exp[-(x^2 + y^2)]$$

$$= \frac{u}{\sqrt{(u^2 - v^2)}} 4v(u^2 - v^2)\, v \exp[-(u^2)]$$

$= 4uv \exp(-u^2)$, $u \geq 0$, $0 \leq v \leq u$

Therefore $f_{UV}(u,v) = 4uv \exp(-u^2)$, $u \geq 0$, $0 \leq v \leq u$
$= 0$, otherwise

NOTES

Now the pdf of U is given by

$$f_U(u) = \int_{-\infty}^{\infty} f_{UV}(u,v)\, dv$$

$$= \int_{0}^{u} 4uv \exp(-u^2) dv = 2u^3 \exp(-u^2)$$

Therefore pdf of U is

$f_U(u) = 2u^3 \exp(-u^2)$, $u \geq 0$
$= 0$, elsewhere

Try yourself !

1.The joint pdf of X and Y is given by $f(x, y) = e^{-(x+y)}$, $x > 0, y > 0$. Find the pdf of $(x+y)/2$.

(Solution: pdf is $4u\, e^{-2u}$, $u \geq 0$)

2.If the pdf of a two dimensional random variable (X, Y) is given by $f(x, y) = x + y$, $0 \leq x, y \leq 1$. Find the pdf of $U = XY$.

(**Solution:** $\frac{e^{-|u|}}{2}$, $-\infty < x, y < \infty$.)

3.If X and Y are independent random variables having density functions

$f_X(x) = 2e^{-2x}$, $x \geq 0$ and $f_Y(y) = 3e^{-3y}$, $y \geq 0$
$0, \quad x < 0$ $0, \quad y < 0$

Find the density functions of their sum $U = X + Y$.
(**solution:** $6e^{-2u}(1 - e^{-u})$, $u > 0$)

REFERENCES:

1. T.Veerarajan, "Probability, statistics and Random Process ", Tata McGraw Hill, 2002.
2. P.Kandasamy, K. Thilagavathi and K. Gunavathi,"Probability, Random Variables and Random processors", S. Chand, 2003.
3. S.C Gupta and V.K Kapoor,"Fundementals of Mathemetical Statistics",Sultan Chand & Sons, 2002

NOTES

UNIT 3
TESTING OF HYPOTHESES

- **Introduction**
- **Tests based on normal distribution (Large sample test)**
- **Student's t test**
- **Senedecor's F test**
- **$\chi 2$ test**

3.1 INTRODUCTION

For the purpose of determining population characteristics, instead of enumerating the entire population, the individuals in the individuals in the sample only are observed. Then the sample characteristics are utilized to estimate the characteristics of the population. For example, on examining the sample of a particular stuff we arrive at a decision of purchasing or rejecting that stuff.

Sampling is quite often used in our day-to-day practical life. For example in a shop we assess the quality of the sugar, wheat or any other commodity by taking a handful of it from the bag and then decide to purchase it or not. A housewife normally tests the cooked products to find if they are properly cooked and contain the proper quantity of salt.

3.2 LEARNING OBJECTIVES

The student will be able to:

- List the steps of hypothesis testing.
- State in your own words the type I and type II errors for a given problem.
- Extract the appropriate information from a story problem to perform a complete hypothesis test.
- Set up the null and alternative hypotheses correctly.
- Choose the appropriate test statistic.
- Choose the appropriate level of significance.

NOTES

- Find the critical value using a table and state the decision rule correctly. Make a statistical decision.
- State the conclusion.
- Perform a hypothesis test for 2 means using the appropriate formula.
- Choose when to use a 2-sample t-test vs. a 2-sample z-test.
- List the assumptions for a 2-sample equal (pooled) variance independent test.
- Perform a 2-sample equal (pooled) variance t-test
- If the problem asks for a business decision based on the hypothesis test, state the appropriate decision.
- Use an F-test to perform an equality of variance hypothesis test.
- Incorporate the F-test for equality of variance in the hypothesis test for 2 means.
- Interpret the results of the chi-square test of independence.
- Look up the critical value in the chi-square table.

Generally population refers to a collection of entities such that each entity possesses an attribute called a characteristic. A statistical hypothesis, is a chain either about the value of a single population characteristics or about the values of several population characteristic.

➢ *Population*

A statistical population is the set of all possible measurements on data corresponding to the entire collection of units for which an inference is to be made.

➢ *Parameter and statistic*

You will be knowing how to find arithmetic mean, median, mode, standard deviation etc from the data contained in a sample. These are called some characterizations of a statistical distribution. These characteristics are called ***parameters*** if they are calculated for a population and are called ***statistics*** if they are calculated for a sample.

For example mean of a population is called a parameter and mean of a sample is called a **statistic.**

The values of the statistic will normally vary from one sample to another, as the values of the population members included in different samples, though drawn from the same population , may be different. These differences in the values of the statistic are said to be sampling fluctuations.

➢ *Sampling distribution.*

These statistics vary from sample to sample if repeated random samples of the same size are drawn from a statistical population. The probability distribution of such a statistic is called the ***sampling distribution***.

NOTES

- *Standard error* (S.E)

If a random variable X is normally distributed with mean μ and standard deviation σ then the random variable X (the mean of a simple random sample of size n) is also normally distributed with mean μ and standard deviation

$$\sigma_x = \frac{\sigma}{\sqrt{n}}$$

The standard deviation of the sampling distribution of mean referred to as the **standard error** of the mean and denoted by $\sigma_x = \frac{\sigma}{\sqrt{n}}$

For finite population **standard error** of the mean is given by

$$\sigma_x = \frac{\sigma}{\sqrt{n}} \frac{\sqrt{(N-n)}}{\sqrt{N-1)}}$$

where N is the number of elements in the population and n is the number of elements in the sample.

- *Estimation and Testing of Hypothesis*

In sampling theory, we primarily concerned with two types of problems which are given below:

a) Some characteristic or feature of the population in which we are interested may be completely unknown to us and we may like to make a guess about this characteristic entirely on the basis of a random sample drawn from the population. This type of problem is known as the problem of estimation.
b) Some information regarding the characteristic or feature of the population may be available to us and we may like to know whether the information is acceptable in the light of the random sample drawn from the population and if it can be accepted, with what degree of confidence it can be accepted. This type of problem is known as the problem of ***testing of hypothesis***.

Hypothesis testing addresses the important question of how to choose among alternative propositions while controlling and minimizing the risk of wrong decisions.

When we attempt to make decisions about the population on the basis of sample information , we have to make assumptions about the nature of the population involved or about the value of some parameter of the population. Such assumptions,

which may or may not be true, are called ***statistical hypothesis***.

We set up a hypothesis which assumes that there is no significant difference between the sample statistic and the corresponding population parameter or between two sample statistics. Such a hypothesis of no difference is called a ***null hypothesis*** and is denoted by H_0.

NOTES

A hypothesis complementary to the null hypothesis is called an ***alternative hypothesis*** and is denoted by H_1.

A procedure for deciding whether to accept or to reject a null hypothesis and hence to reject or to accept the alternative hypothesis is called the test of hypothesis.

- *Test of significance*

The difference between θ_0 and θ where θ_0 is a parameter of the population and θ is the corresponding sample statistic, which is caused due to sampling fluctuations is called ***insignificant difference.***

The difference that arises due to the reason that either the sampling procedure is not purely random or that the sample has not been drawn from the given population is known as ***significant difference***.

This procedure of testing whether the difference between θ_0 and θ is significant or not is called as the ***test of significance***.

- *Critical region*

The critical region of a test of statistical hypothesis is that the region of the normal curve which corresponds to the rejection of null hypothesis.

- *Level of significance*

Level of significance is the probability level below which the null hypothesis is rejected. Generally, 5% and 1% level of significance are used.

- *Errors in hypothesis*

The level of significance is fixed by the investigator and as such it may be fixed at a higher level by his wrong judgment. Due to this, the region of rejection becomes larger and the probability of rejecting a null hypothesis, when it is true, becomes greater. The error committed in rejecting H_0, when it is really true, is called ***Type I error.***

This is similar to a good product being rejected by the consumer and hence Type I error is also known as *producer's risk*.

The error committed in accepting H_0, when it is false, is called ***Type II error***. As this error is similar to that of accepting a product of inferior quality, it is also known as *consumer's risk.*

The probabilities of committing Type I and II errors are denoted by α & β respectively. It is to be noted that the probability of α of committing Type I error is the ***level of significance.***

NOTES

➢ *One Tailed and two tailed tests*

If θ_0 is a parameter of the population and θ is the corresponding sample statistic and if we set up the null hypothesis $H_0: \theta = \theta_0$, then the alternative hypothesis which is complementary to H_0 can be anyone of the following:

i) $H_1: \theta \neq \theta_0$, i.e., $\theta > \theta_0$ or $\theta < \theta_0$
ii) $H_1: \theta > \theta_0$
iii) $H_1: \theta < \theta_0$

H_1 given in (i) is called a two tailed alternative hypothesis, whereas H_1 given in (ii) is called a right-tailed alternative hypothesis and H_1 given in (iii) is called a left-tailed alternative hypothesis.

When H_0 is tested while H_1 is a one-tailed alternative (right or left), the test of hypothesis is called a one-tailed test.

When H_0 is tested while H_1 is a two-tailed alternative (right or left), the test of hypothesis is called a two-tailed test.

➢ *Critical values or significant values*

The value of test statistic which separates the critical (or rejection) region and the acceptance region is called the critical value or significant value. It depends upon:
i)the level of significance used and
ii)the alternative hypothesis, whether it is two tails or single tailed.

The critical value of the test statistic a level of significance α for a two tailed test is given by z_α where z_α is determined by the equation
$P(|Z| > z_\alpha) = \alpha$
i.e., z_α is the value so that the total area of the critical region on both tails is α. Since normal probability curve is a symmetrical curve, we get

$$P(Z > z_\alpha) + P(Z < -z_\alpha) = \alpha$$

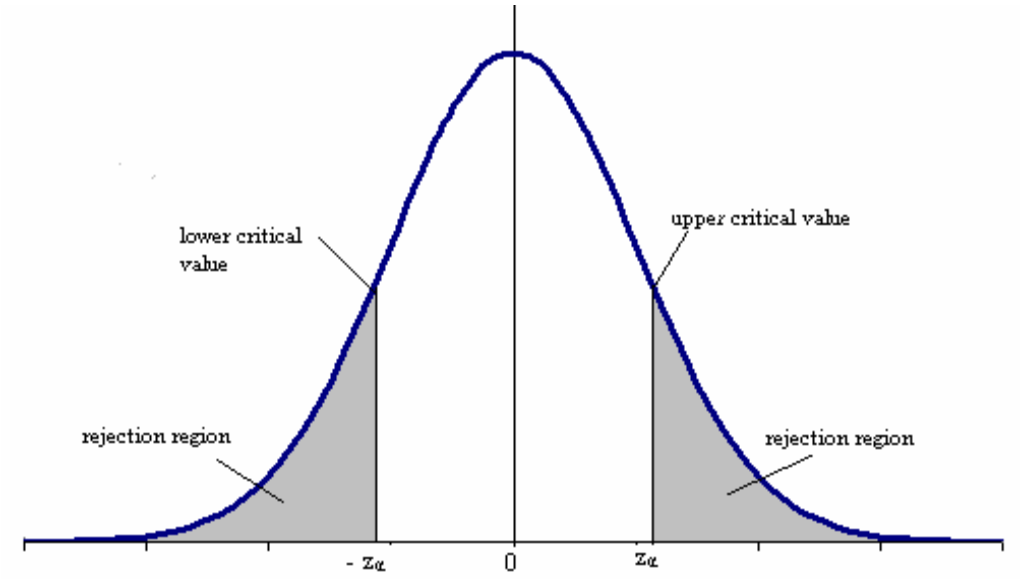
$$\Rightarrow P(Z > z_\alpha) + P(Z > z_\alpha) = \alpha$$

$$\Rightarrow 2P(Z > z_\alpha) = \alpha$$

$$\Rightarrow P(Z > z_\alpha) = \alpha/2$$

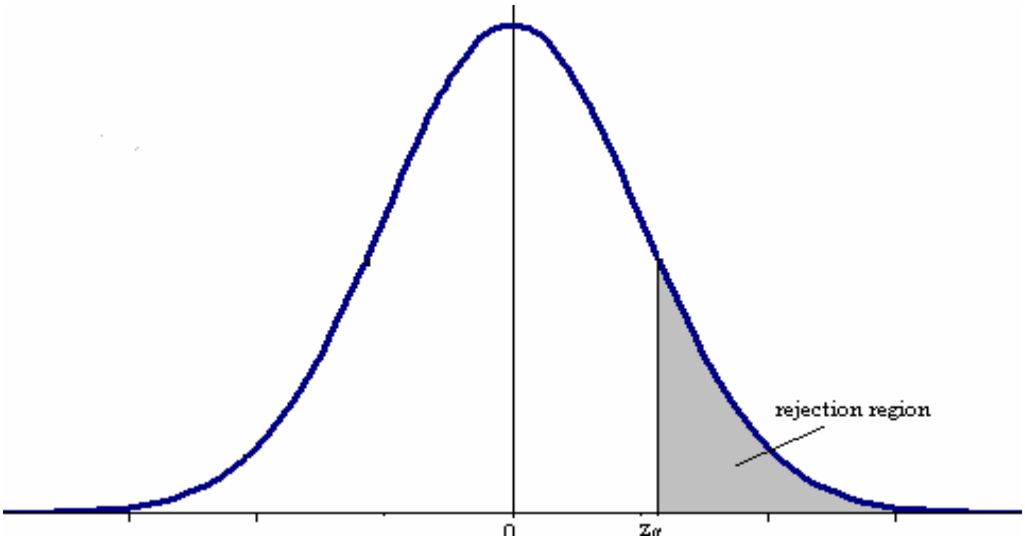
⇨ i.e., the area of each tail is $\alpha/2$. Thus z_α is the value such that area to the right of z_{α} is $\alpha/2$ and to the left $-z_\alpha$ is $\alpha/2$.

NOTES



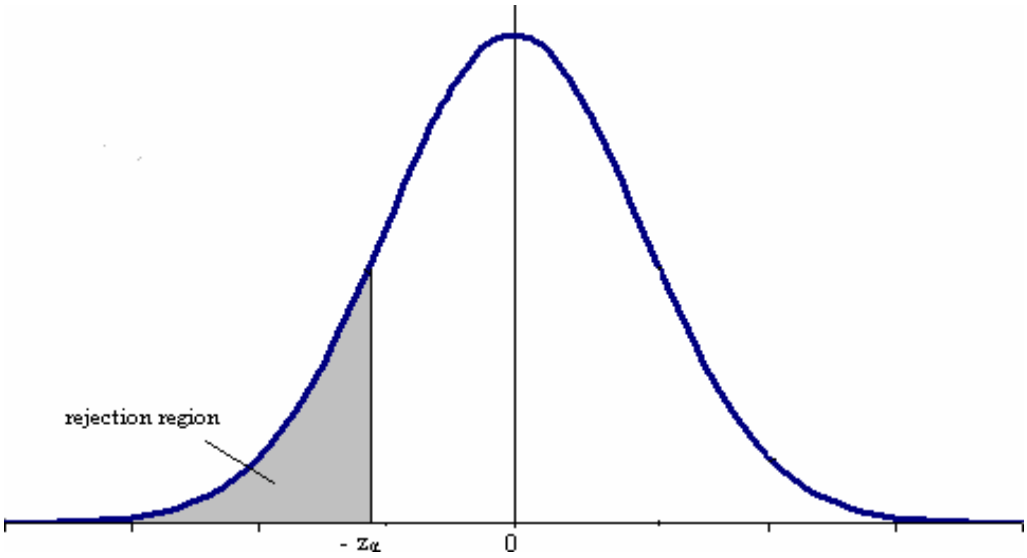
TWO-TAILED TEST
(Level of significance α)

In case of single tail alternative, the critical value z_α is determined so that total area to the right of it is α and for left-tailed test the total area to the left is $- z_\alpha$ is z_α, i.e.,
For right tailed test : $P(Z > z_\alpha) = \alpha$
For left tailed test : $P(Z < - z_\alpha) = \alpha$



RIGHT-TAILED TEST
(Level of significance α)

NOTES



LEFT-TAILED TEST
(Level of significance α)

Thus the significant or critical value of Z for a single value of Z for a single-tailed test at level of significance α is same as the critical value of Z for a two tailed test at level of significance '2 α'

The critical values z_α for some standard Level of significance's are given in the following table.

Nature of test	LOS	1%(.01)	2%(.02)	5%(.05)	10%(.1)
Two-tailed		$\lvert z_\alpha \rvert = 2.58$	$\lvert z_\alpha \rvert = 2.33$	$\lvert z_\alpha \rvert = 1.96$	$\lvert z_\alpha \rvert = 1.645$
Right-tailed		$z\alpha = 2.33$	$z_\alpha = 2.055$	$z_\alpha = 1.645$	$z_\alpha = 1.28$
Left-tailed		$z_\alpha = -2.33$	$z_\alpha = -2.055$	$z_\alpha = -1.645$	$z_\alpha = -1.28$

Procedure for testing of hypothesis

1. Null Hypothesis H_0 is defined.
2. Alternative hypothesis H_1 is also defined after a careful study of the problem and also the nature of the test(whether one Tailed or two tailed tests) is decided.
3. LOS(Level of significance) 'α' is fixed or taken from the problem if specified and z_α is noted.
4. The test-statistic $z = \dfrac{X - E(X)}{S.E\,(X)}$ is computed
5. Comparison is made between $|z| > z_\alpha$, H_0 is rejected or H_1 is accepted, i.e., it is concluded that the difference between x and E(x) is significant at α LOS.

NOTES

➢ *Confidence or Fiducial limits and Confidence interval*

Confidence interval is an interval that provides lower and upper limits for a specific population parameter is expected to lie. The two values of the statistic which determine the limits of the interval are called confidence limits. Thus confidence interval is the interval in which a population parameter is expected to lie with certain probability.
For example 95% confidence interval for population mean μ is

$$[\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}}]$$

3.3 TEST BASED ON NORMAL DISTRIBUTION

Tests of significance of large samples

It is generally agreed that, if the size of the sample exceeds 30, it should be regarded as a large sample. The tests of significance used for large samples are different from the ones used for small samples for the reason that the following assumptions made for large sample do not hold for small samples.

1. The sampling distribution of a statistic is approximately normal, irrespective of whether the distribution of the population is normal or not.

2. Sample statistics are sufficiently close to the corresponding population parameters and hence may be used to calculate the standard error of the sampling distribution.

3.3.1 TEST 1

Test of significance of the difference between sample mean and population mean.

Let $X_1, X_2, \ldots X_n$ be the sample observations in a sample of size n, drawn from a population that is $N(\mu, \sigma)$

Then each X_i follows $N(\mu, \sigma)$. Then their mean $\bar{X}$ follows a $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$

Even if the population, from which the sample is drawn, is non-normal, it is known that the above result holds good, provided n is large.

Therefore the test statistic $z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$

As usual , if $|z| \leq z_\alpha$, the difference between the sample mean X and the population mean μ is not significant at α% LOS.

NOTES

Note:

1. If σ is not known, the sample S.D. 's' can be used in its place, as s is nearly equal to σ when n is large.

2. 95 % confidence limits for μ are given by $\frac{|\overline{X} - \mu|}{\sigma/\sqrt{n}} = 1.96$

i.e., $[\overline{X} - 1.96 \frac{\sigma}{\sqrt{n}}, \overline{X} + 1.96 \frac{\sigma}{\sqrt{n}}]$, if σ is known. If σ is not known, then the 95 % confidence interval is ., $[\overline{X} - 1.96 \frac{s}{\sqrt{n}}, \overline{X} + 1.96 \frac{s}{\sqrt{n}}]$

3.3.2 TEST 2

Test of significance of the difference between the means of two samples.

Let $\overline{X}_1$ and $\overline{X}_2$ be the means of two large samples of sizes n_1 and n_2 drawn from two populations (normal or non-normal) with the same mean μ and variances σ_1^2 and σ_2^2 respectively.

Then $\overline{X}_1$ follows a $N\left(\mu, \frac{\sigma_1}{\sqrt{n_1}}\right)$ and X_2 follows a $N\left(\mu, \frac{\sigma_2}{\sqrt{n_2}}\right)$ either exactly or approximately.

Therefore $\overline{X}_1$ and $\overline{X}_2$ follows a normal distribution.

$E(\overline{X}_1 - \overline{X}_2) = E(\overline{X}_1) - E(\overline{X}_2) = \mu - \mu = 0$

$V(\overline{X}_1 - \overline{X}_2) = V(\overline{X}_1) + V(\overline{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$

(since $\overline{X}_1$ & $\overline{X}_2$ are independent, as the samples are independent)

Thus $(X_1 - X_2)$ follows a $N\left\{0, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right\}$

Therefore the test statistic $z = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$

If $|z| \leq z_\alpha$, the difference $(\overline{X}_1 - \overline{X}_2)$ and 0 or the difference between $\overline{X}_1$ and $\overline{X}_2$ is not significant at α % LOS.

Note:

1. If the samples are drawn from the same population, i.e., if $\sigma_1 = \sigma_2 = \sigma$ then

NOTES

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

2.If σ_1 and σ_2 are not known and $\sigma_1 \neq \sigma_2$, σ_1 and σ_2 can be approximated by the sample S.D's s_1 and s_2 .

Therefore the test statistic $z = \dfrac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ (a)

3.If σ_1 and σ_2 are equal and not known, then $\sigma_1 = \sigma_2 = \sigma$ is approximated by

$\sigma^2 = \dfrac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2}$. Hence in such a situation ,

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Therefore the test statistic $z = \dfrac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_2} + \frac{s_2^2}{n_1}}}$ (b)

The difference in the denominators of the values of z is given in (a) and (b) may be noted.

Example 1: A random sample of 200 tins of coconut oil gave an average weight of 4.95 kg with a standard deviation of 0.21 kg. Do we accept the hypothesis of net weight 5 kg per tin at 5% level ?

Solution:

Sample size, n = 200
Sample mean $\bar{x}$ = 4.95 kg
Sample SD s = 0.21 kg
Population mean μ = 5 kg
The sample is a large sample and so we apply z-test

$H_0 : \bar{x} = \mu$

$H_1 : \bar{x} \neq \mu$

NOTES

The test statistic $z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$

(when σ is not known, replace σ by s)

Therefore the test statistic $z = \frac{\bar{x} - \mu}{s/\sqrt{n}}$

$= \frac{4\ .95\ -\ 5}{0.21/\sqrt{200}} = -3.37$

Therefore $|z| = 3.37$

At 1% level of significance the tabulated value of z is 2.58

H_0 is rejected at 1% level since the calculated value of |z| is greater than the table value of z. Therefore the net weight of a tin is not equal to 5 kg.

Example 2: A sample of 900 items has mean 3.4 and standard deviation 2.61. Can the sample be regarded as drawn from a population with mean 3.25 at 5% level of significance?

Solution:

Sample size, n = 900

Sample mean x = 3.4

Sample SD s = 2.61

Population mean $\mu = 3.25$

The sample is a large sample and so we apply z-test

$H_0 : \bar{x} = \mu$

$H_1 : \bar{x} \neq \mu$

The test statistic $z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$

(when σ is not known, replace σ by s)

Therefore the test statistic $z = \frac{\bar{x} - \mu}{s/\sqrt{n}}$

$= \frac{3.4 - 3.25}{2.61/\sqrt{900}} = 1.72$

Therefore $|z| = 1.72$

At 1% level of significance the tabulated value of z is 2.58

H_0 is accepted since the calculated value is less than the table value. Therefore it is likely that the sample belongs to the population with mean 3.4

NOTES

Example 3: The mean breaking strength of the cables supplied by a manufacturer is 1800 with a SD of 100.By a new technique in the manufacturing process, it is claimed that the breaking strength of the cable has increased. In order to test this claim, a sample of 50 cables is tested and it is found that the mean breaking strength is 1850. Can we support the claim at 1% level of significance.

Solution:

Sample size, n = 50

Sample mean $\bar{x}$ = 1850

Population SD σ = 100

Population mean μ = 1800

The sample is a large sample and so we apply z-test

$H_0 : \bar{x} = \mu$

$H_1 : \bar{x} > \mu$ (one-tailed test)

The test statistic $z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$

$= \dfrac{1850 - 1800}{100/\sqrt{50}} = 3.54$

Therefore |z| = 1.72

At 1% level of significance the tabulated value of z is 2.33

H_0 is rejected and H_1 is accepted at 1% level since the calculated value of |z| is greater than the table value of z. i.e., is based on the sample data, we may support the claim of increase in breaking strength.

Example 4: The mean value of a random sample of 60 items was found to be 145 with a SD of 40. Find the 95% confidence limits for the population mean. What size of the Sample is required to estimate the population mean within five of its actual value with 95% or more confidence, using the sample mean?

Solution:

Sample size, n = 60

Sample mean $\bar{x}$ = 145

Sample SD s = 40

95 % confidence limits for μ are given by $\left| \dfrac{x - \mu}{\sigma/\sqrt{n}} \right| \leq 1.96$

NOTES

i.e., $[\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}}]$, if σ is known. If σ is not known, then the 95 % confidence interval is ., $[\bar{x} - 1.96 \frac{s}{\sqrt{n}}, \bar{x} + 1.96 \frac{s}{\sqrt{n}}]$

i.e., $[\bar{x} - 1.96 \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96 \frac{s}{\sqrt{n}}]$

i.e., $145 - \frac{1.96 \times 40}{\sqrt{60}} \leq \mu \leq 145 + \frac{1.96 \times 40}{\sqrt{60}}$

i.e., $134.9 \leq \mu \leq 155.1$

We have to find the value of n such that $P\{\bar{x} - 5 = \mu \leq \bar{x} + 5\} \geq 0.95$

$P\{-5 \leq \mu - \bar{x} \leq 5\} = 0.95$

$P\{|\mu - \bar{x}| \leq 5\} \geq 0.95$

$P\{|\bar{x} - \mu| \leq 5\} \geq 0.95$

$P\{\frac{|\bar{x} - \mu|}{\sigma/\sqrt{n}} \leq \frac{5}{\sigma/\sqrt{n}}\} \geq 0.95$

$P\{|z| \leq \frac{5\sqrt{n}}{\sigma}\} \geq 0.95$ (1)

where z is the standard normal variate.

We know that $P\{|z| \leq 1.96\} = 0.95$

Therefore the least value of $n = n_L$ that will satisfy (1) is given by $\frac{5\sqrt{n_L}}{\sigma} = 1.96$

i.e., $\sqrt{n_L} = \frac{1.96\,\sigma}{5} = \frac{1.96\,s}{5}$

$n_L = \left[\frac{1.96 \times 40}{5}\right]^2 = 245.86$

Therefore the least size of the sample = 246

Example 5: A normal population has mean of 0.1 and SD of 2.1. Find the probability that the mean of a samples of size 900 drawn from this population will be negative.

Solution:

Since $\bar{x}$ follows a $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$, $z = \frac{|\bar{x} - \mu|}{\sigma/\sqrt{n}}$ is the standard normal variate.

NOTES

Now $P(\bar{x} < 0) = P\{\bar{x} - 0.1 < -0.1\}$

$$= P\left\{ \frac{\bar{x} - 0.1}{(2.1)/\sqrt{900}} < \frac{-0.1}{(2.1)/\sqrt{900}} \right\}$$

$$= P\{ z < -1.43\}$$

$$= P\{ z > 1.43\} \text{ (by symmetry of the standard normal distribution)}$$

$$= 0.5 - P\{ 0 < z < 1.43\}$$

$$= 0.5 - 0.4236 \text{ (from the normal table)}$$

$$P(\bar{x} < 0) = 0.0764$$

Example 6: A college conducts both day and night classes intended to be identical. A sample of 100 day class students, yields examination results as mean 72.4 and SD 14.8 and a sample of 200 class students yields examination results as mean 73.9 and SD 17.9. Are the two means statistically equal at 10% level.

Solution:

$\bar{x}_1 = 72.4$ $\bar{x}_2 = 73.9$
$s_1 = 14.8$ $s_2 = 17.91$
$n_1 = 100$ $n_2 = 200$

The two given samples are large samples

$H_0 : \mu_1 = \mu_2$ or $\bar{x}_1 = \bar{x}_2$

$H_1 : \mu_1 \neq \mu_2$ or $\bar{x}_1 \neq \bar{x}_2$

The test statistic

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$= \frac{72.4 - 73.9}{\sqrt{[(14.8)^2/100 + (17.9)^2/200]}} = -0.77$$

$|z| = 0.77$
The table value of z at 10% level = 1.645
H_0 is accepted at 10% level, since the calculated value is less than the table value.
Therefore two means are statistically equal.

NOTES

Example 7: The sales manager of a large company conducted a sample survey in states A and B taking 400 samples in each case. The results were –

	State A	State B
Average Sales	Rs.2,500.00	Rs.2,200.00
SD	Rs.400.00	Rs.550.00

Test whether the average sales is the same in the two states at 1% level.

Solution:

$\bar{x}_1 = 2000 \quad \bar{x}_2 = 2200$
$s_1 = 400 \quad s_2 = 550$
$n_1 = 400 \quad n_2 = 400$

The two given samples are large samples

$H_0 : \mu_1 = \mu_2$ or $\bar{x}_1 = \bar{x}_2$

$H_1 : \mu_1 \neq \mu_2$ or $\bar{x}_1 \neq \bar{x}_2$

The test statistic $$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$= \frac{2500 - 2200}{\sqrt{[\,(400)^2/400 + (550)^2/400\,]}} = 8.82$$

$|z| = 8.82$
The table value of z at 1% level = 2.58

The calculated value of z is greater than the table value of z.
Therefore H_0 is rejected at 1% level , i.e., the average sales within two states differ significantly.

Example 8: In a random sample of size 500, the mean is found to be 20. In another independent sample of size 400, the mean is 15. Could the samples have been drawn from the same population with SD at 4.

Solution:

$\bar{x}_1 = 20 \quad \bar{x}_2 = 15$
$n_1 = 500 \quad n_2 = 400$

NOTES

$\sigma = 4$

The two given samples are large samples

$H_0 : \bar{x}_1 = \bar{x}_2$

$H_1 : \bar{x}_1 \neq \bar{x}_2$ (two-tailed test)

The test statistic $$z = \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$= \frac{20 - 15}{4\sqrt{[\,1/500 + 1/400\,]}} = 18.6$$

$|z| = 18.6$

The table value of z at 1% level = 2.58

The calculated value of z is greater than the table value of z.

Therefore H_0 is rejected at 1% level , i.e., the sample could not have been drawn from the same population.

Example 9: Test the significance of the difference between the means of the samples, drawn from two normal populations with same SD from the following data:

Sample 1	Size	Mean	SD
	100	61	4
Sample 2	200	63	6

Solution:

$H_0 : \mu_1 = \mu_2$ or $\bar{x}_1 = \bar{x}_2$

$H_1 : \mu_1 \neq \mu_2$ or $\bar{x}_1 \neq \bar{x}_2$

The population have same SD.

The test statistic $$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_2} + \frac{s_2^2}{n_1}}}$$

$$= \frac{61 - 63}{\sqrt{[\,(4)^2/200 + (6)^2/100\,]}} = -3.02$$

NOTES

$|z| = 3.02$
The table value of z at 5% level = 1.96

The calculated value of z is greater than the table value of z.

Therefore H_0 is rejected at 5% level , i.e., the two normal populations, from which the samples are drawn , may not have the same mean, though they may have the same SD.

Example 10: The mean height of 50 male students who showed above average participation in college athletics was 68.2 inches with a SD of 2.5 inches, while 50 male students who showed no interest in such participation has a mean height of 67.5 inches with a SD of 2.8 inches. Test the hypothesis that male students who participated in college athletics are taller than other male students.

Solution:
Athletic Non-Athletic

$\bar{x}_1 = 68.2$" $\bar{x}_2 = 67.5$"
$s_1 = 2.5$" $s_2 = 2.8$"
$n_1 = 50$ $n_2 = 50$

The two given samples are large samples

$H_0 : \mu_1 = \mu_2$ or $\bar{x}_1 = \bar{x}_2$

$H_1 : \mu_1 > \mu_2$ or $\bar{x}_1 > \bar{x}_2$ (one-tailed test)

The test statistic $$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$= \frac{68.2 - 67.5}{\sqrt{[\,(2.5)^2/50 + (2.8)^2/50\,]}} = 1.32$$

$|z| = 1.32$

The table value of z at 5% level = 1.645
The calculated value of z is less than the table value of z.
Therefore H_0 is accepted and H_1 is rejected at 5% level .
Therefore we cannot say that athletics are taller than non-athletics.

Example 11: The average marks scored by 32 boys is 72 with a SD of 8 while that of 36 girls is 70 with a SD of 6. Test at 1% of significance whether the boys perform better than the girls.

NOTES

Solution:

$\overline{x_1} = 72$ $\overline{x_2} = 70$
$s_1 = 8$ $s_2 = 6$
$n_1 = 32$ $n_2 = 36$

The two given samples are large samples

$H_0 : \mu_1 = \mu_2$ or $\overline{x}_1 = \overline{x}_2$

$H_1 : \mu_1 > \mu_2$ or $\overline{x}_1 > \overline{x}_2$ (one-tailed test)

The test statistic $$z = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$= \frac{72 - 70}{\sqrt{[\,(8)^2/32 + (6)^2/36\,]}} = 1.15$$

$|z| = 1.15$

The table value of z at 1% level = 2.33

The calculated value of z is less than the table value of z.
Therefore H_0 is accepted and H_1 is rejected at 5% level .

Therefore we cannot say that boys perform better than girls.

Example 12: The heights of men in a city are normally distributed with mean 171 cm and SD 7 cm. While the corresponding value for women in the same city are 165 cm and 6 cm respectively. If a man and a women are chosen at random from this city, find the probability that the women is taller than the man.

Solution:
Let $\overline{x}_1$ and $\overline{x}_2$ denote the mean heights of men and women respectively.
Then $\overline{x}_1$ follows N(171, 7) and $\overline{x}_2$ follows a N(165, 6).

$\overline{x}_1 - \overline{x}_2$ also follows normal distribution.

$E(\overline{x}_1 - \overline{x}_2) = E(\overline{x}_1) - E(\overline{x}_2) = 171 - 165 = 6$

$V(\overline{x}_1 - \overline{x}_2) = V(\overline{x}_1) + V(\overline{x}_2) = 49 + 36 = 85$

S.D of $(\overline{x}_1 - \overline{x}_2) = \sqrt{85} = 9.22$

NOTES

S.D of $(\bar{x}_1 - \bar{x}_2)$ follows a N(6, 9.22)

Now $P(\bar{x}_2 > \bar{x}_1) = P(\bar{x}_1 - \bar{x}_2 < 0)$

$$= P\left\{\frac{(\bar{x}_1 - \bar{x}_2) - 6}{9.22} < \frac{-6}{9.22}\right\}$$

$= P\{z < -0.65\}$, where z is the standard normal variate.

$= P\{z > 0.65\}$ by symmetry.

$= 0.5 - P(0 < z < 0.65)$

$= 0.5 - 0.2422 = 0.2578.$

3.3.3 TEST 3

Test of significance of the difference between sample proportion and population proportion.

Let X be the number of successes in n independent Bernoulli trial in which the probability of success for each trial is a constant = p(say). Then it is known that X follows a binomial distribution with mean E(X) = np and variance V(X) = nPQ.

When n is large, X follows N(np, $\sqrt{(nPQ)}$), i.e., a normal distribution with mean nP and S.D, $\sqrt{(nPQ)}$, where Q = 1 – P.

$\frac{X}{n}$ follows $N\left\{\frac{np}{n}, \sqrt{\frac{(nPQ)}{n^2}}\right\}$

Now $\frac{X}{n}$ is the proportion of success in the sample consisting of n trials, that is denoted by p. Thus the sample proportion p follows $N\left\{P, \sqrt{\frac{PQ}{1}}\right\}$

Therefore test statistic $z = \frac{p - P}{\sqrt{(PQ/n)}}$

If $|z| \geq z_\alpha$, the difference between the sample proportion p and the population mean P is not significant at α% LOS.

Note:

1. If P is not known, we assume that p is nearly equal to P and hence S.E.(p) is taken as $\sqrt{(pq/n)}$. Thus $z = \frac{p - P}{\sqrt{(pq/n)}}$

NOTES

2. 95% confidence limits for P are then given by $\frac{|P - p|}{\sqrt{(pq/n)}} \leq 1.96$, i.e. they are

$[\, p - 1.96\sqrt{(pq/n)} \;,\; p + 1.96\sqrt{(pq/n)} \,]$

3.3.4 TEST 4

Test of significance difference between two sample proportions.

Let p_1 and p_2 be the proportions of successes in two large samples of size n_1 and n_2 respectively drawn from the same population or from two population with same proportion P.

Then p_1 follows $N\left\{ P, \sqrt{\frac{PQ}{n_1}} \right\}$

and p_2 follows $N\left\{ P, \sqrt{\frac{PQ}{n_2}} \right\}$

Therefore $p_1 - p_2$, which is a linear combination of two normal variables also follows normal distribution.

Now $E(p_1 - p_2) = E(p_1) - E(p_2) = p - p = 0$

$V(p_1 - p_2) = V(p_1) + V(p_2)$ (since two samples are independent)

$$= PQ\left[\frac{1}{n_1} + \frac{1}{n_2}\right]$$

Therefore $(p_1 - p_2)$ follows $N\left\{ 0, \sqrt{PQ\left[\frac{1}{n_1} + \frac{1}{n_2}\right]} \right\}$

Therefore the test statistic $z = \dfrac{(p_1 - p_2)}{\sqrt{PQ\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}}$

If P is not known, an unbiased estimate of P based on both samples, given by $\frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$ is used in the place of P.

If $|z| \leq z_\alpha$, the difference between the two sample proportion p_1 and p_2 is not significant at α% LOS.

Example 13: While throwing 5 dice 30 times, a person obtains success 23 times (securing a 6 was considered a success). Can we consider the difference between the observed and the expected results as being significantly different.

NOTES

Solution:
Sample size = 5 × 30 = 150
Sample proportion , p = 23/150
Population proportion, P = 1/6 (Q = 1 – P = 5/6)

H_0: P = 1/6
H_1: P ≠ 1/6

The test statistic $z = \frac{p - P}{\sqrt{(PQ/n)}} = \frac{23/150 - 1/6}{\sqrt{[1/6 \times 5/6 /150]}} = -0.438$

Therefore |z| = 0.438

The table value of z at 5% level = 1.96

The calculated value of z is less than the table value of z.
Therefore H_0 is accepted at 5% level .

Therefore the difference between the sample proportion and the population proportion is not significant.

Example 14: In a certain city, 380 men out of 800 are found to be smokers. Discuss whether this information supports the view that majority of the men in this city are non smokers.

Solution:

Sample size = 800

Sample proportion of non-smokers, p = 420 / 800

Population proportion, P = 1/2 (Q = 1 – P = ½)

H_0: P = 1/2
H_1: P > 1/6 (majority of men are non-smokers: one tail test)

The test statistic $z = \frac{p - P}{\sqrt{(PQ/n)}} = \frac{420 / 800 - 1 / 2}{\sqrt{[1/2 \times 1/2 /800]}} = 1.414$

Therefore z = 1.414

The table value of z for one tail test at 5% level = 1.645

The calculated value of z is less than the table value of z.

NOTES

Therefore H_0 is accepted and H_1 is rejected at 5% level .
Therefore we cannot conclude that majority are non-smokers.

Example 15: Experience has shown that 20% of a manufactured product is of top quality. In one days production of 400 articles, only 50 are of top quality. Show that either the production of the day chosen was not representative sample or the hypothesis of 20% was wrong. Based on the particulars days production, find also the 95% confidence levels for the % of top quality products.

Solution:
Sample size = 400
Sample proportion of non-smokers, p = 50/ 400 = 1/8
Population proportion, P = (20%) = 1/5 (Q = 1 – P = 4/5)

H_0: P = 1/5
H_1: P ≠ 1/5

The test statistic $z = \frac{p - P}{\sqrt{(PQ/n)}} = \frac{1/8 - 1/5}{\sqrt{[1/5 \times 4/5 /400]}} = -3.75$

Therefore |z| = 3.75
The table value of z at 5% level = 1.96
The calculated value of z is greater than the table value of z.
Therefore H_0 is rejected at 5% level .

Therefore the production of the particular day chosen is not a representative sample.

95% confidence limits for P are then given by $\frac{|P - p|}{\sqrt{(pq/n)}} \leq 1.96$,

We have taken $\sqrt{(pq/n)}$ in the denominator, because P is assumed to be unknown , For which we are trying to find the confidence limits and P is nearly equal to p.

i.e. $[p - 1.96\sqrt{(pq/n)} \leq P \leq p + 1.96\sqrt{(pq/n)}]$
i.e., $0.125 - 1.96 \times \sqrt{[1/8 \times 7/8 /400]} \leq P \leq 0.125 + 1.96 \times \sqrt{[1/8 \times 7/8 /400]}$
i.e., $0.093 \leq P \leq 0.157$
Therefore 95% confidence limits for the percentage of top quality product are 9.3 and 15.7.

Example 16: Show that for a random sample of size 100 drawn with replacement the standard error of sample proportion cannot exceed 0.05

Solution:

The items of the sample are drawn one after another replacement.
Therefore the proportion of success in the population, i.e., P remains a constant.

We know that the sample proportion p follows N P, $\frac{PQ}{1}$

i.e., standard error of p = v(pq/n) = $\frac{1}{10}$ v(pq) (n = 100) (1)

Now ($\sqrt{P}$ - $\sqrt{Q}$)^2 $\geq$ 0
P + Q - 2vPQ $\geq$ 0
1 - 2$\sqrt{PQ}$ $\geq$ 0
or $\sqrt{PQ} \leq \frac{1}{2}$ (2)

using (2) in (1), we get
S E of p $\leq$ 1/20. that is standard error of p cannot exceed 0.05.

Example 17: A cubicle die is thrown 9000 times and the throw of 3 or 4 is observed 3240 times. Show that the die cannot be regarded as an unbiased one.

Solution:

H_0: the die is unbiased, i.e., P = 1/3 (= the probability of getting 3 or 4)
H_1 : P $\neq$ 1/3 (two tailed test)

Though we may test the significance of difference between the sample and population proportions, we shall test the significance of the difference between the number X of successes in the sample and that in the population.

When n is large, X follows N(np, $\sqrt{(nPQ)}$), i.e., a normal distribution with mean nP and S.D, v(nPQ), where Q = 1 – P.

Therefore $z = \frac{X - np}{\sqrt{(nPQ)}}$

$= \frac{3240 - (9000 \times 1/3)}{\sqrt{[9000 \times 1/3 \times 2/3]}} = 5.37$

|z| = 5.37
The table value of z at 5% level = 1.96

The calculated value of z is greater than the table value of z.
Therefore H_0 is rejected at 5% level .
Therefore the die cannot be regarded as unbiased.

NOTES

NOTES

Example 18: A company has its head office at Calcutta and a branch at Mumbai. The personal Director wanted to know if the workers in the two places would like the introduction of a new plan of work and a survey was conducted for this purpose. Out of a sample of 500 workers at Calcutta, 62% favored the new plan. At Mumbai, out of a sample of 400 workers, 41% were against the plan. Is there any significant difference between the 2 groups in their attitude towards the new plan at 5% level.

Solution:

$n_1 = 500$ $n_2 = 400$
$p_1 = 62/100$ $p_2 = 59/100$

$H_0 : P_1 = P_2$ (proportions in the two places are equal)
$H_0 : P_1 \neq P_2$

The test statistic $$z = \frac{(p_1 - p_2)}{\sqrt{PQ\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}}$$

If P is not known, an unbiased estimate of P based on both samples, given by $\frac{n_1p_1 + n_2p_2}{n_1 + n_2}$ is used in the place of P.

$$P = \frac{500 \times 62/100 + 400 \times 59/100}{500 + 400} = 0.607 : \quad Q = 0.393$$

$$\text{Therefore } z = \frac{0.62 - 0.59}{\sqrt{[0.607 \times 0.393\ (1/500 + 1/400)]}} = 0.9146$$

The table value of z at 5% level = 1.96

The calculated value of z is less than the table value of z.
Therefore H_0 is accepted at 5% level .

Therefore there is no significant difference in their attitude towards the introduction of new plan.

Example 19: Before increase in excise duty of tea, 400 people out of a sample of 500 persons were found to be tea drinkers. After an increase in duty 400 people were tea drinkers out of a sample of 600 people. Using the standard error of proportion state whether there is a significant decrease in the consumption of tea.

NOTES

Solution:

$n_1 = 500$ $n_2 = 60$

$p_1 = 400/100$ $p_2 = 400/600$

$H_0 : P_1 = P_2$

$H_0 : P_1 > P_2$ (one tail test)

The test statistic $z = \dfrac{(p_1 - p_2)}{\sqrt{PQ\left[\dfrac{1}{n_1} + \dfrac{1}{n_2}\right]}}$

If P is not known, an unbiased estimate of P based on both samples, given by $\dfrac{n_1p_1 + n_2p_{2,}}{n_1 + n_2}$ is used in the place of P.

$P = \dfrac{500 \times 400/500 + 600 \times 400/600}{500 + 600} = 8/11 : \quad Q = 3/11$

Therefore $z = \dfrac{400/500 - 400/600.}{\sqrt{[8/11 \times 3/11 (1/500 + 1/600)]}} = 4.81$

The table value of z at 1% level for a one-tail test = 2.33

The calculated value of z is greater than the table value of z.
Therefore H_0 is rejected and H_1 is accepted at 1% level .

Therefore there is a significant decrease in the consumption of tea after the increase in the excise duty.

Example 20: 15.5 % of a random sample of 1600 under-graduates were smokers. Whereas 20% of a random sample of 900 post graduates were smokers in a state. Can we conclude that less number of under graduates are smokers than the post graduates.

Solution:

$n_1 = 1600$ $n_2 = 900$

$p_1 = 0.155$ $p_2 = 0.2$

$H_0 : P_1 = P_2$

$H_0 : P_1 < P_2$ (one tail test)

The test statistic $z = \dfrac{(p_1 - p_2)}{\sqrt{PQ\left[\dfrac{1}{n_1} + \dfrac{1}{n_2}\right]}}$

NOTES

If P is not known, an unbiased estimate of P based on both samples, given by $\frac{n_1p_1 + n_2p_2,}{n_1 + n_2}$ is used in the place of P.

$$P = \frac{1600 \times 0.155 + 900 \times 0.2}{1600 + 900} = 0.1712$$

$$\text{Therefore } z = \frac{0.155 - 0.2}{\sqrt{[0.1712 \times 0.8288 \times (1/1600 + 1/900)]}} = -2.87$$

The table value of z at 5% level for a one-tail (left tailed) test = -1.645

$|z| > |z_\alpha|$

The calculated value of z is greater than the table value of z.
Therefore H_0 is rejected and H_1 is accepted at 5% level .

Therefore the habit of smoking is less among the undergraduates than among the postgraduates .

3.3.5 TEST 5

Test of significance of the difference between sample S.D and population S.D

Let 's' be the S.D of a large sample of size n drawn from a normal population with S.D σ. Then it is known that s follows a $N[\sigma, \sigma/\sqrt{(2n)}]$ approximately.

Then the test statistic $z = \frac{s - \sigma}{\sigma/\sqrt{(2n)}}$

As before the significance of the difference between s and σ is tested.

3.3.6 TEST 6

Test of significance of the difference between sample S.D's of two large samples.

Let s_1 and s_2 be the S.D's of two large samples of sizes n_1 and n_2 drawn from a normal population with S.D Test of significance of the difference between sample S.D and population S.D σ.

s_1 follows a $N[\sigma, \sigma/\sqrt{(2n_1)}]$ and s_2 follows a $N[\sigma, \sigma/\sqrt{(2n_2)}]$

Therefore $(s_1 - s_2)$ follows $N\left\{0, \sigma\sqrt{\frac{1}{2n_1} + \frac{1}{2n_2}}\right\}$

NOTES

Therefore the test statistic $z = \left\{ \frac{s_1 - s_2}{\sigma\sqrt{\frac{1}{2n_1} + \frac{1}{2n_2}}} \right\}$

As usual, the significance of the difference between s_1 and s_2 is tested.

Note:

If σ is not known, it is approximated by $= \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2}}$.

where n_1 and n_2 are large . In this situation

the test statistic $z = \frac{s_1 - s_2}{\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2} \cdot \left[\frac{1}{2n_1} + \frac{1}{2n_2}\right]}}$

$z = \frac{s_1 - s_2}{\sqrt{\frac{s_1^2}{2n_2} + \frac{s_2^2}{2n_1}}}$

Example 21: A manufacturer of electric bulbs according to a certain process finds the SD of life of the lamps to be 100 hours. He wants to change the process if the new process results in a smaller variation in the life of the lamps. In adopting the new process a sample of 150 bulbs gave an SD of 95 hours. Is the manufacturer justified in changing the process.

Solution:

$\sigma = 100$, $n = 150$ and $s = 95$
$H_0 : s = \sigma$
$H_0 : s < \sigma$ (left tailed test)

Then the test statistic $z = \frac{s - \sigma}{\sigma/\sqrt{(2n)}}$

$= \frac{95 - 100}{100 / \sqrt{300}} = -0.866$

The table value of z at 5% level (left tailed) = -1.645
$|z| < |z_\alpha|$

The calculated value of z is less than the table value of z.
Therefore H_0 is accepted and H_1 is rejected at 5% level .

Hence the manufacturer is not justified in changing the process.

NOTES

Example 22: In two random samples of sizes of 150 and 250 the SD were calculated as 15.3 and 13.8. Can we conclude that the samples are drawn from the populations with the same SD.

Solution:

$n_1 = 150$ $n_2 = 250$
$s_1 = 15.3$ $s_2 = 13.8$

$H_0 : \sigma_1 = \sigma_2$ (The sample belong to the populations with same standard deviation)
$H_0 : \sigma_1 \neq \sigma_2$ (The sample belong to the populations with different standard deviation)

If σ is not known,
the test statistic $z = \dfrac{s_1 - s_2}{\sqrt{\dfrac{s_1^2}{2n_2} + \dfrac{s_2^2}{2n_1}}}$

$$= \frac{15.3 - 13.8}{\sqrt{[\,(15.3)^2/300 - (13.8)^2/500\,]}}.$$

$$= 1.5 / 1.0770 = 1.39$$

The table value of z at 5% level = 1.96

The calculated value of z is less than the table value of z.
Therefore H_0 is accepted at 5% level .

Hence the sample belong to the populations with the same SD.

How you understood ?

1. Define sampling distribution and standard error. Obtain standard error of mean when population is large.
2. What is mean t by statistical hypothesis? What are the two types of errors of decision that arise in testing a hypotheses ?
3. Define null hypotheses and alternative hypotheses ?
4. What do you mean by critical region and acceptance region ?
5. What is the relation between critical values for a single tailed and two-tailed test.

TRY YOURSELF !

1) A sample of 400 male students is found to have a mean height 171.38 cm. Can it be reasonably regarded as sample from large population with mean height of 171.17 cm and standard deviation 3.3 cm?

NOTES

2) An automatic machine fills in tea in sealed tins with mean weight of tea 1 kg and SD 1 gm. A random sample of 50 tins was examined and it was found that their mean weight was 999.50 gm. Is the machine working properly?

3) Two random samples of sizes 400 and 500 have mean 10.9 and 11.5 respectively. Can the sample be regarded as drawn from the same population with variance 25?

4) A person buys 100 electric tubes from well known makes taken at random from stocks for testing purpose. HE finds that 'make A' has a mean life of 1300 hours with a SD of 82 hours and 'make B' has mean life of 1248 hours with a SD of 93 hours. Discuss the significance of these results to test which make of electric tube should the person buy?

5) A person threw 10 dice 500 times and obtained 2560 times 4,5 or 6. Can this be attributed to fluctuation in sampling?

6) A manufacturer claimed that at least 95% of the equipment which he supplied to a factory confirmed to specifications. An examination of a sample of 200 pieces of equipment revealed that 18 were faulty. Test his claim at a significance level of 0.01, 0.05.

7) A coin was tossed 900 times and head appeared 490 times. Does the result support the hypothesis that the coin I unbiased?

8) In a sample of 600 men from a certain city, 450 men are found to be smokers. In a sample of 900 from another city 450 are found to be smokers. Do the data indicate that the two cities are significantly different with respect to prevalence of smoking habit among men?

9) A machine produced 20 defective articles in a batch of 400. After overhauling it produced 10 defectives in a batch of 300. Has the machine improved?

10) In a year there are 956 births in a town A of which 52.5% were males whereas in town A and B combined this proportion in a total of 1406 births was 0.496. Is there any significant difference in the proportion of male and female birth in the two towns?
(here $n_1 = 956$, $n_2 = 450$, $p_1 = 502/956 = 0.525$, $p_2 = 192/450 = 0.427$, $P = 0.496$)

11) The standard deviation of a sample of size 50 is 63. Could this have come from a normal population with standard deviation6?

NOTES

3.4 STUDENT'S T- DISTRIBUTION

Tests of significance for small samples.

When the sample is small, i.e., $n < 30$, the sampling distributions of many statistics are not normal, even though the parent populations may be normal. Therefore the tests of significance discussed in the previous section are not suitable for small tests. Consequently we have to develop entirely different tests of significance that are applicable to small samples.

Student's t- distribution

A random variable T is said to follow student' t- distribution or simply t-distribution, if its probability density function is given by

$$f(t) = \frac{1}{\sqrt{\nu}\,\beta(\nu/2, 1/2)}\left[1 + \frac{t^2}{\nu}\right]^{-(\nu+1)/2}, \quad -\infty < t < \infty$$

ν is called the number of degrees of freedom of the t-distribution.

Note on degrees of freedom (d.f.):
The number of independent variates which make up the statistic is known as the degrees of freedom and is usually denoted by ν (the letter 'Nu' of the Greek alphabet.)

The number of degrees of freedom, in general, is the total number of observations less the number of independent constraints imposed on the observations.

t- distribution was defined by the mathematician W.S.D Gosset whose pen name is student, hence the name student's-t distribution.

❖ Uses of t-test

This statistic is used in the following situations in tests of hypothesis.

i) It is used to test whether a specific value is the population mean when the given sample is a small sample and the population S.D is not known.

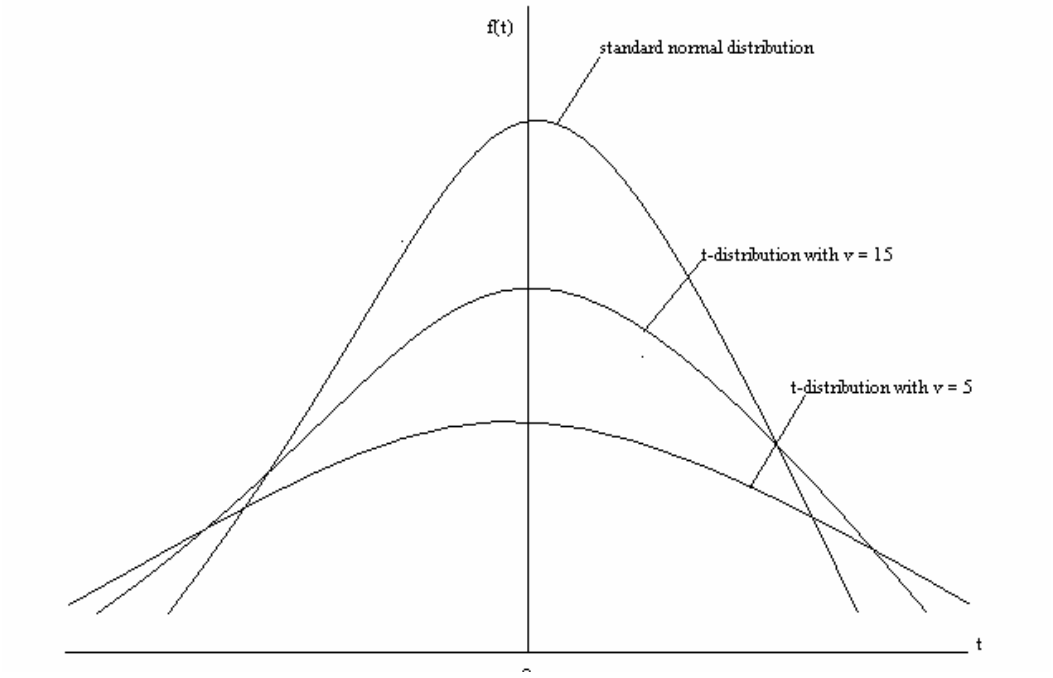
ii) It is also used to test the significance of difference between the means of two populations based on to small samples of sizes n_1 and n_2 when the S.D's of the population are not known and also the samples drawn are independent.

iii) It is also used to test the significance of difference between the means of paired observations.

NOTES

❖ Properties of the sampling distribution of t.

i) The probability curve of the t-distribution is similar to the standard normal curve and symmetric about t =0, bell shaped and asymptotic to the t-axis.
ii) It has greater dispersion than the normal distribution.
iii) It has uni-modal distribution.
iv) The shape of the curve varies as the number of degrees of freedom varies.
v) For sufficiently large values of n, the t-distribution tends to the standard normal distribution.
vi) The mean of t-distribution is zero.



❖ Critical Values of t and the t-table.

The critical value of t at level of significance **α** and degrees of freedom *v* is given by $P\{|t| > t_v(\alpha)\} = \alpha$ for two tailed test, as in the case of normal distribution and large samples and by $P\{t > t_v(\alpha)\} = \alpha$ for the right-tailed test also, as in the case of normal distribution .The critical value of t for a single (right or left) tailed test at LOS 'α' corresponding to *v* degrees of freedom is the same as that for a two-tailed test at LOS '2 α' corresponding to the same degrees of freedom.

Critical values $t_v(\alpha)$ of the t-distribution for two-tailed tests corresponding to a few important levels of significance and a range of values of *v* have been published by Prof. R.A.Fisher in the form of a table, called the t-table.

NOTES

3.4.1 TEST 1

Test of significance of the difference between sample mean and population mean.

If x is the mean of a sample of size n, and s is the sample standard deviation the test statistic is given by

$$t = \frac{\bar{x} - \mu}{s / \sqrt{(n-1)}}$$

This t-statistic follows a t-distribution with number of degrees of freedom $\nu = n - 1$.

Sometimes $t = \frac{\bar{x} - \mu}{S / \sqrt{n}}$

where $S^2 = \sum_{r=1}^{n} \frac{(\bar{x}_i - \bar{x})^2}{n-1}$ and is called student's-t.

We shall use only $t = \frac{\bar{x} - \mu}{s / \sqrt{(n-1)}}$, where s is the sample S.D.

We get the value of $t_\nu(\alpha)$ for the LOS α and $\nu = n - 1$.from the table.

If the calculated value of t satisfies $|t| > t_\nu(\alpha)$, the null hypothesis H_0 is accepted at LOS 'α' otherwise, H_0 is rejected at LOS α.

Note:

95% confidence interval of m is given by

$$= \frac{\bar{x} - \mu}{s / \sqrt{(n-1)}} \leq t_{0.05}, \text{ since}$$

$$P\left\{\left|\frac{\bar{x} - \mu}{s / \sqrt{(n-1)}}\right|\right\} \leq t_{0.05} = 0.95$$

i.e., by $\bar{x} - t_{0.05}\frac{s}{\sqrt{(n-1}} \leq m \leq \bar{x} + t_{0.05}\frac{s}{\sqrt{(n-1}}$ where $t_{0.05}$ is the 5 % critical value of for n – 1 degrees of freedom for a two tailed test.

3.4.2 TEST 2

Test of significance of the difference between means of two small samples drawn from the same normal population.

The test statistic is given by $t = \frac{\bar{x}_1 - \bar{x}_2}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

NOTES

where $S^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$

$n_1 + n_2 - 2$ is the number of degrees of freedom of the statistic.

Note:

If $n_1 = n_2 = n$ and if the pairs of values of X_1 and X_2 are associated in some way or correlated we shall assume that H_0: $\bar{d}$ $(= \bar{x} - \bar{y}) = 0$ and test the significance of the difference between $\bar{d}$ and 0, using the test statistic $t = \frac{\bar{d}}{s / \sqrt{(n-1)}}$. with $\nu = n - 1$,

where $d_i = x_i - y_i$ (i = 1,2, …n), $\bar{d} = \bar{x} - \bar{y}$: and

s = S.D of d's = $\frac{1}{n} \sum_{i=1}^{n} (d_i - d)^2$

Example 1: A sample of 10 house owners is drawn and the following values of their incomes are obtained. Mean Rs.6,000.00 ; SD Rs. 650.00. test the hypothesis that the average income of house owners of the town is Rs.5,500.00.

Solution:

$n = 10 \quad s = 650$

$\bar{x} = 6{,}000 \quad \mu_0 = 5{,}500$

since the sample size $n = 10 < 30$, the sample is a small sample. Therefore we have to apply t-test for testing the mean.

H_0: $\bar{x} = \mu$ (ie the average income of the house owners of the town is Rs.5,500)
H_1: $\bar{x} \neq \mu$

If $\bar{x}$ is the mean of a sample of size n, and s is the sample standard deviation the test statistic is given by

$$t = \frac{\bar{x} - \mu}{s / \sqrt{(n-1)}}$$

This t-statistic follows a t-distribution with number of degrees of freedom $\nu = n - 1$.

$$t = \frac{6{,}000 - 5{,}500}{650 / \sqrt{9}} = 2.31$$

Number of degrees of freedom = n – 1 = 9
The table value of t for 9 degrees of freedom at 5% level = 2.262

H_0 is rejected since the calculated value of t > the table value of t. Hence the average income of house owners in that tow is not Rs.5,500/-

NOTES

Example 2: A machine is designed to produce insulating washers for electrical devices of average thickness of 0.025 cm. A random sample of 10 washers was found to have an average thickness of 0.024 cm with a standard deviation of 0.002 cm. Test the significance of deviation.

Solution:

$n = 10$ $s = 0.002$

$\bar{x} = 0.024$ cm $\mu = 0.025$cm

since the sample size $n = 10 < 30$, the sample is a small sample. Therefore we have to apply t-test for testing the mean.

$H_0: \bar{x} = \mu$
$H_1: \bar{x} \neq \mu$

If x is the mean of a sample of size n, and s is the sample standard deviation the test statistic is given by

$$t = \frac{\bar{x} - \mu}{s / \sqrt{(n-1)}}$$

This t-statistic follows a t-distribution with number of degrees of freedom $v = n - 1$.

$$t = \frac{0.024 - 0.025}{0.002 / \sqrt{9}} = -1.5$$

$|t| = 1.5$
Number of degrees of freedom $= n - 1 = 9$
The table value of t for 9 degrees of freedom at 5% level = 2.262

H_0 is accepted since the calculated value of $|t| <$ the table value of t. Hence deviation is not significant.

Example 3: The mean lifetime of 25 bulbs is found as 1550 hours with a SD of 120 hours. The company manufacturing the bulbs claims that the average life of their bulbs is 1600 hours. Is this claim acceptable at 5% level of significance?

Solution:

$n = 25$ $s = 120$

$\bar{x} = 1550$ $\mu = 1600$

since the sample size $n = 10 < 30$, the sample is a small sample. Therefore we have to apply t-test for testing the mean.

NOTES

H_0: $\bar{x} = \mu$
H_1: $\bar{x} < \mu$ (left tailed test)

If $\bar{x}$ is the mean of a sample of size n, and s is the sample standard deviation the test statistic is given by

$$t = \frac{\bar{x} - \mu}{s / \sqrt{(n-1)}}.$$

This t-statistic follows a t-distribution with number of degrees of freedom $v = n - 1$.

$$t = \frac{1550 - 1600}{120 / \sqrt{24}} = -2.04$$

$|t| = 2.04$

Number of degrees of freedom = n – 1 = 24

The table value of t for 24 degrees of freedom at 5% level for one-tailed test =
= The table value of t for 24 degrees of freedom at 10% level for two-tailed test
= 1.71

H_0 is rejected and H_1 is accepted since the calculated value of |t| > the table value of t. Therefore the claim of company cannot be accepted at 5% LOS.

Example 4: A filling machine is expected to fill 5 kg of powder into bags. A sample of 10 bags gave the weights 4.7, 4.9, 5.0, 5.1, 5.4, 5.2, 4.6, 5.1, 4.6 and 4.7. test whether the machine is working properly.

Solution:

n = 10 μ = 5 kg

Let us calculate x and s from the sample data

x	x^2
4.7	22.09
4.9	24.01
5.0	25.00
5.1	26.01
5.4	29.16
5.2	27.04
4.6	21.16
5.1	26.01
4.6	21.16
4.7	22.09
49.3	243.73

NOTES

$\overline{x} = \Sigma x / n = 49.3/ 10 = 4.93$

$s^2 = \frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2$

$s^2 = 243.73/10 - (4.93)^2$

$s = \sqrt{0.073} = 0.27$

since the sample size n = 10 < 30, the sample is a small sample. Therefore we have to apply t-test for testing the mean.

$H_0: \overline{x} = \mu$
$H_1: \overline{x} \neq \mu$

If x is the mean of a sample of size n, and s is the sample standard deviation the test statistic is given by

$$t = \frac{\overline{x} - \mu}{s / \sqrt{(n-1)}}$$

This t-statistic follows a t-distribution with number of degrees of freedom $v = n - 1$.

$$t = \frac{4.93 - 5}{0.27 / \sqrt{9}} = -0.78$$

$|t| = 0.78$
Number of degrees of freedom = n – 1 = 9

The table value of t for 9 degrees of freedom at 5% level for one-tailed test = 2.262

H_0 is accepted since the calculated value of |t| < the table value of t.
Hence the machine is working properly.

Example 5: The heights of 10 males of a given locality are found to be 175, 168, 155, 170, 152, 170, 175, 160 and 165 cms. Based on this sample of 10 items, test the hypothesis that the mean height of males is 170 cms. Also find the 95% confidence levels for the height of the males in that locality.

Solution:

n = 10 $\mu = 170$

NOTES

Let us calculate x and s from the sample data

x	d	d^2
175	10	100
168	3	9
155	-10	100
170	5	25
152	-13	169
170	5	25
175	10	100
160	-5	25
160	-5	25
165	0	0
1650	0	578

$\bar{x} = \Sigma x / n = 1650/ 10 = 165$

$s^2 = \frac{\Sigma d^2}{n} - \left(\frac{\Sigma d}{n}\right)^2$

$s^2 = 578/10 - 0$

$s = \sqrt{57.8} = 7.6$

since the sample size n = 10 < 30, the sample is a small sample. Therefore we have to apply t-test for testing the mean.

$H_0: \bar{x} = \mu$

$H_1: \bar{x} \neq \mu$

If $\bar{x}$ is the mean of a sample of size n, and s is the sample standard deviation the test statistic is given by

$$t = \frac{\bar{x} - \mu}{s / \sqrt{(n-1)}}.$$

This t-statistic follows a t-distribution with number of degrees of freedom $v = n - 1$.

$$t = \frac{165 - 170}{7.6/ \sqrt{9}} = -1.97$$

$| t | = 1.97$

Number of degrees of freedom = n – 1 = 9

The table value of t for 9 degrees of freedom at 5% level for = 2.26

NOTES

H_0 is accepted since the calculated value of |t| < the table value of t.
This means the mean height of males can be regarded as 170 cm.

95% confidence interval of m is given by

$$= \frac{\bar{x} - \mu}{s / \sqrt{(n-1)}} \leq t_{0.05},$$

i.e., by $[\bar{x} - t_{0.05} \frac{s}{\sqrt{n-1}} \leq m \leq \bar{x} + t_{0.05} \frac{s}{\sqrt{n-1}}]$ where $t_{0.05}$ is the 5 % critical value of t for n – 1 degrees of freedom for a two tailed test.

$$165 - 2.26 \frac{7.6}{\sqrt{9}} \leq m \leq 165 + 2.26 \frac{7.6}{\sqrt{9}}$$

$$159.3 \leq m \leq 170.7$$

i.e., the heights of males in the locality are likely to lie within 159.3 cm and 170.7 cm.

Example 6 : A certain injection administered to each 12 patients resulted in the following increases of blood pressure : 5, 2, 8, -1, 3, 0, 6, -2, 1, 5, 0, 4. Can it be concluded that the injection will be, in general, accompanied by an increase in B.P?.

Solution:

n = 12 μ = 1600

Let us calculate x and s from the sample data

x	x^2
5	25
2	4
8	64
-1	1
3	9
0	0
6	36
-2	4
1	1
5	25
0	0
4	16
31	185

NOTES

$\overline{x} = \Sigma x / n = 31 / 12 = 2.58$

$s^2 = \frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2$

$s^2 = 185/12 - (2.58)^2$

$s = \sqrt{2.96}$

since the sample size n = 10 < 30, the sample is a small sample. Therefore we have to apply t-test for testing the mean.

H_0: $\overline{x} = \mu$ (where $\mu = 0$ i.e., the injection will not result in increase in B.P)
H_1: $\overline{x} > \mu$ (Right tailed test)

If x is the mean of a sample of size n, and s is the sample standard deviation the test statistic is given by

$$t = \frac{\overline{x} - \mu}{s / \sqrt{(n-1)}}.$$

This t-statistic follows a t-distribution with number of degrees of freedom $v = n - 1$.

$$t = \frac{2.58 - 0}{2.96 / \sqrt{11}} = 2.89$$

$|t| = 2.89$

Number of degrees of freedom = n – 1 = 11
The table value of t for 11 degrees of freedom at 5% level for one-tailed test =
= The table value of t for 11 degrees of freedom at 10% level for two-tailed test
= 1.80

H_0 is accepted and H_1 is accepted since the calculated value of |t| < the table value of t. i.e., we may conclude that the injection is accompanied by an increase in B.P.

Example 7: Two samples of 6 and 5 items respectively gave the following data :
Mean of the 1st sample = 40
SD of the 1st sample = 8
Mean of the 2st sample = 50
SD of the 2st sample = 10
Is the difference between the means significant? The value of t for 9df at 5% level is 2.26.

Solution:
The two given samples are small samples. Let us apply t-test for testing the mean.

NOTES

$H_0 : \mu_1 = \mu_2$ (The means of the two population are equal)
$H_1 : \mu_1 ? \mu_2$ (The means of the two population are not equal)
$n_1 = 6 \qquad n_2 = 5$

$\bar{x}_1 = 40 \qquad \bar{x}_2 = 50$

$s_1 = 8 \qquad s_2 = 10$

The test statistic is given by $t = \dfrac{\bar{x}_1 - \bar{x}_2}{S\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$

where $S^2 = \dfrac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$

$n_1 + n_2 - 2$ is the number of degrees of freedom of the statistic.

$S^2 = \dfrac{6 \times 64 + 5 \times 100}{6 + 5 - 2} = 98.22$

$S = 9.91$

Therefore $t = \dfrac{40 - 50}{9.91\sqrt{(1/6 + 1/5)}} = -1.65$

$|t| = 1.65$

number of degrees of freedom, ndf = 6 + 5 – 2 = 9

The table value of t for 9 df at 5% level = 2.262

The calculated value of t < the table value of t.
H_0 is accepted at 5% level. Hence there is no significant difference between the means of the population.

Example 8: Below are given the gains in weights (lbs) of cows fed on two diets X and Y. Gain in weight (in lbs)

Diet X	25	32	30	32	24	14	32			
Diet Y	24	34	22	30	42	31	40	30	32	35

Test at 5% level, whether the two diets differ as regards their effect on mean increase in weight (table value of t for 15df at 5% is 2.131)

Solution:
The two given samples are small samples. Let us apply t-test for testing the mean.

NOTES

$H_0 : \mu_1 = \mu_2$ (The means of the two population are equal)
$H_1 : \mu_1 \neq \mu_2$ (The means of the two population are not equal)
Let us calculate the mean and S.D of the two samples –

x	d_1 (x-27)	d_1^2	y	d_2 (y-32)	d_2^2
25	-2	4	24	-8	64
32	5	25	34	2	4
30	3	9	22	-10	100
32	5	25	30	-2	4
24	3	9	42	10	100
14	-5	169	31	-1	1
32		25	40	8	64
			30	-2	4
			32	0	0
			35	3	9
189			266	320	350

$n_1 = 7 \qquad n_2 = 10$

$\bar{x} = \Sigma x / n_1 = 189 / 7 = 27$

$\bar{y} = \Sigma y / n_2 = 320 / 10 = 32$

Let $d_1 = x - 27$ $d_2 = y - 32$

$$s_1^2 = \frac{\Sigma d_1^2}{n_1} - \left(\frac{\Sigma d_1}{n_1}\right)^2$$

$s_1^2 = 266/7 - 0 = 38$

$$s_2^2 = \frac{\Sigma d_2^2}{n_2} - \left(\frac{\Sigma d_2}{n_2}\right)^2$$

$s_2^2 = 350/10 = 35$

The test statistic is given by $t = \dfrac{\bar{x}_1 - \bar{x}_2}{S\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$

where $S^2 = \dfrac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$

$n_1 + n_2 - 2$ is the number of degrees of freedom of the statistic.

NOTES

$S^2 = \frac{7 \times 38 + 10 \times 35}{7 + 10 - 2} = 41.07$

$S = 6.41$

Therefore $t = \frac{27 - 32}{6.41 \sqrt{(1/7 + 1/10)}} = -1.59$

$|t| = 1.59$

number of degrees of freedom, ndf = 7 + 10– 2 = 15

The table value of t for 15 df at 5% level = 2.131

The calculated value of t < the table value of t.

H_0 is accepted at 5% level. Hence there is no significant difference between the mean increase in the weight due to two diets.

Example 9 : A group of 5 patients treated with medicine A weigh 42, 39, 48, 60 and 41 kgs; A second group of 7 patients from the same hospital treated with medicine B weigh 38, 42, 56, 64, 68, 69 and 62 kgs. Do you agree with the claim that the medicine B increases weight significantly. (the value of t at 5% significance for 10df is 2.228)

Solution:

The two given samples are small samples. Let us apply t-test for testing the mean.

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 < \mu_2$ (Medicine B increases significantly) (one tailed test)

$n_1 = 5 \quad n_2 = 7$

x	d_1 (x-46)	d_1^2	y	d_2 (y-57)	d_2^2
42	-4	16	38	-19	361
39	-7	49	42	-15	225
48	2	4	56	-1	1
60	14	196	64	7	49
41	-5	25	68	11	121
			69	12	144
230	0	290	399	0	926

$\bar{x} = \Sigma x / n_1 = 230 / 5 = 46$

$\bar{y} = \Sigma y / n_2 = 399 / 7 = 57$

NOTES

Let $d_1 = x - 46$ $d_2 = y - 57$

$$s_1^2 = \frac{\Sigma d_1^2}{n_1} - \left(\frac{\Sigma d_1}{n_1}\right)^2$$

$s_1^2 = 290 / 5 - 0 = 290 / 5$

$$s_2^2 = \frac{\Sigma d_2^2}{n_2} - \left(\frac{\Sigma d_2}{n_2}\right)^2$$

$s_2^2 = 926 / 7 - 0 = 926 / 7$

The test statistic is given by $t = \dfrac{x_1 - x_2}{S\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$

where $S^2 = \dfrac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$

$n_1 + n_2 - 2$ is the number of degrees of freedom of the statistic.

$S^2 = \dfrac{5 \times 58 + 10 \times 926 / 7}{5 + 7 - 2} = 121.6$

$S = 11.03$

Therefore $t = \dfrac{46 - 57}{11.03\sqrt{(1/5 + 1/7)}} = -1.7$

$|t| = 1.7$

number of degrees of freedom, ndf = $5 + 7 - 2 = 10$

The table value of t for 10 df at 5% level for one tailed test =
The table value of t for 10 df at 10% level for two tailed test = 1.812

The calculated value of t < the table value of t.

H_0 is accepted at 5% level.
Therefore medicine A and B do not differ significantly w.r.t increase in weights.

Example 10 : The marks obtained by a group of 9 regular course students and another group of 11 part time course students in a test are given below –

Regular	56	62	63	54	60	51	67	69	58		
Part time	62	70	71	62	60	56	75	64	72	68	66

NOTES

Examine whether the marks obtained by regular students and part time students differ significantly at 5% level of significance and 1% level of significance.

Solution:

The two given samples are small samples. Let us apply t-test for testing the mean.

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$

Let us calculate the mean and variance of the two samples.

x	d_1	d_1^2	y	d_2 (y-57)	d_2^2
56	-4	16	62	-4	16
62	2	4	70	4	16
63	3	9	71	5	25
54	-6	36	62	-4	16
60	0	0	60	-6	36
51	-9	81	56	-10	100
67	7	49	75	9	81
69	9	81	64	-2	4
58	-2	4	72	6	36
			68	2	4
			66	0	0
540	0	280	726	0	334

$n_1 = 9 \qquad n_2 = 11$

$\bar{x} = \Sigma x / n_1 = 230 / 5 = 46$

$\bar{y} = \Sigma y / n_2 = 399 / 7 = 57$

Let $d_1 = x - 46 \quad d_2 = y - 57$

$$s_1^2 = \frac{\Sigma d_1^2}{n_1} - \frac{\Sigma d_1}{n_1}^2$$

$s_1^2 = 280 / 9 - 0 = 280 / 9$

NOTES

$s_2^2 = \frac{\Sigma d_2^2}{n_2} - \left(\frac{\Sigma d_2}{n_2}\right)^2$

$s_2^2 = 334/11 - 0 = 334/11$

The test statistic is given by $t = \frac{x_1 - x_2}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

where $S^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$

$n_1 + n_2 - 2$ is the number of degrees of freedom of the statistic.

$S^2 = \frac{9 \times 280/9 + 11 \times 334/11}{9 + 11 - 2} = 34.11$

$S = 5.84$

Therefore $t = \frac{60 - 66}{5.84\sqrt{(1/9 + 1/11)}} = -2.28$

$|t| = 2.28$

number of degrees of freedom, ndf = $9 + 11 - 2 = 18$

The table value of t for 18 df at 5% level = 2.101
The calculated value of t >the table value of t.

H_0 is rejected at 5% level.
Therefore the marks obtained by regular students and part-time students differ significantly.

Example 11 : The following data relate to the marks obtained by 11 students in two tests, one held at the beginning of the year and the other at the end of the year after intensive coaching. Do the data indicate that the students have benefited by coaching?

Test 1	19	23	16	24	17	18	20	18	21	19	20
Test 2	17	24	20	24	20	22	20	20	18	22	19

Solution:
Let $d = x_1 - x_2$, where x_1 & x_2 are the marks in the two tests.

NOTES

Test 1 = x_1	Test 2 = x_2	d = x_1 - x_2	d^2
19	17	2	4
23	24	-1	1
16	20	-4	16
24	24	0	0
17	20	-3	9
18	22	-4	16
20	20	0	0
18	20	-2	4
21	18	3	9
19	22	-3	9
20	19	1	1
		-11	69

$\Sigma d = -11 \quad \Sigma d^2 = 69$

$\bar{d} = \frac{\Sigma d}{n} = -11/11 = -1$

$s^2 = s_d^2 = \frac{\Sigma d^2}{n} - \left(\frac{\Sigma d}{n}\right)^2 = 69/11 - (-1)^2 = 5.27$

$s = 2.296$

If $n_1 = n_2 = n$ and if the pairs of values of X_1 and X_2 are associated in some way or correlated we shall assume that $H_0: \bar{d} (= \bar{x} - \bar{y}) = 0$ and test the significance of the difference between $\bar{d}$ and 0, using the test statistic $t = \frac{\bar{d}}{s / \sqrt{(n-1)}}$. with $v = n - 1$,

where $d_i = x_i - y_i$ (i = 1,2, …n), $\bar{d} = \bar{x} - \bar{y}$: and

s = S.D of d's = $\frac{1}{n} \sum_{i=1}^{n} (d_i - d)^2$

$H_0: \bar{d} = 0$ i.e, the students have not benefited by coaching.

$H_1: \bar{d} < 0$ (i.e., $x_1 < x_2$) one tailed test.

$t = \frac{-1}{2.296 / \sqrt{(10-1)}} = -1.38$, $v = 11 - 1 = 10$

$|t| < 1.38$

NOTES

The table value of t for 10 df at 5% level for one tailed test
= The table value of t for 10 df at 10% level for two tailed test
= 1.81.
The calculated value of t < the table value of t.

H_0 is accepted and H_1 is rejected at 5% level.
Therefore there is no significant difference between the two sets of marks.
i.e., the students have not benefited by coaching.

How you understood ?

1. Write down the probability density of student's t-distribution.
2. State the important properties of the t-distribution.
3. Give any two uses of t-distribution.
4. What do you mean by degrees of freedom ?
5. What is the test statistic used to the significance of the difference between the means of two small sample.

TRY YOURSELF !

1) Certain refined edible oil is packed in tins holding 16 kg each. The filling machine can maintain this but with a standard deviation of 0.5 kg. Samples of 25 are taken from the production line. If a sample mean is i) 16.35 kg ii) 15.8 kg , can we be 95% sure that the sample has come from a population of 16 kg tins?

2) A company has been producing steel tubes of mea inner diameter of 2.00 cm. A sample of 10 tubes gives an inner diameter of 2.01 cm and a variance of 0.0004 cm^2. Is the difference in the value of mean significant?
(Value of t for 9df at 5% level = 2.262)

3) A random sample of 10 boys has the following IQ's: 70, 120, 110, 101, 88, 83, 95, 98, 107, 100. Do these data support the assumption of a population mean IQ of 100 ?

4) A fertilizer mixing machine is set to give 12 kg of nitrate for every quintal bag of fertilizer. Ten 100 kg bags are examined and percentage of nitrate is s follows 11, 14, 13, 12, 13, 12, 13, 14, 11, 12. Is there reason to believe that the machine is defective?

5) Two salesman A and B are working in a certain district. From a sample survey conducted by the head office, the following results were obtained. State whether there is significant difference in the average sales between the two salesmen:

NOTES

	A	B
No: of sales	20	18
Average sales (in Rs)	170	205
Standard Deviation	20	25

6) Two batches of the same product are tested for their mean life. Assuming that he life of the product follow a normal distribution with an unknown variance, test the hypothesis that the mean life is the same for both the batches, gives the following information:

Batch	Sample size	Mean life (in hrs)	SD (in hrs)
I	10	750	12
II	8	820	14

7) Two sets of 10 students selected at random from a college were taken : one set was given memory test as they were and the other was given the memory test after two weeks of training and the scores are given below:

Set A	10	8	7	9	8	10	9	6	7	8
Set B	12	8	8	10	8	11	9	8	9	9

Due you think there is any significant effect due to training?

8) Wire cable is manufactured by two processors. Laboratory tests were performed by putting samples of cables under tension and recording the load required (coded units) to break the cable giving the following data.

Process I	9	4	10	7	9	10
Process II	149	13	12	13	8	10

Can we say that the two processes have the same effect on the mean breaking strength, at 5% level of significance

9) A company is testing two machines. A random sample of 8 employees is selected and each employee uses each machine for one hour. The number of components produced is shown in the following table.

Employee	1	2	3	4	5	6	7	8
I Machine	96	107	84	99	102	87	93	101
II Machine	99	112	90	97	108	97	94	98

NOTES

Test whether there is evidence of difference between the machines in the mean number of components produced.

3.5 VARIANCE RATIO TEST OR F-TEST

This test is used to test the significance of two or more sample estimates of population variance.

The -statistic is defined as a ratio of unbiased estimates of population variance.
Symbolically, $F = \frac{S_1^2}{S_2^2}$

where $S_1^2 = \frac{\Sigma (x_1 - \bar{x}_1)^2}{n_1 - 1}$ and $S_2^2 = \frac{\Sigma (x_2 - \bar{x}_2)^2}{n_2 - 1}$

Therefore the distribution $F = \frac{S_1^2}{S_2^2}$ $(S_1^2 < S_2^2)$ is given by the following pdf

$$f(F) = \frac{1}{B(v_1/2, v_2/2)} \frac{(v_1/v_2)^{v1/2} \; F^{(v1/2 - 1)}}{(1 + v_1 F/v_2)^{(v1 + v2)/2}}. \quad F > 0$$

This is called the distribution of the variance ratio F or Senedecor's F-distribution with $v_1 = n_1 - 1$ and $v_2 = n_2 - 1$ degrees of freedom.

3.5.1 F-test of significance of the difference between population variances and F table.

If s_1^2 and s_2^2 are the variances of two samples of sizes n_1 and n_2 respectively, the estimates of the population variances based on these samples are respectively
$S_1^2 = \frac{n_1 s_1^2}{n_1 - 1}$ and $S_2^2 = \frac{n_2 s_2^2}{n_2 - 1}$

The quantities $v_1 = n_1 - 1$ and $v_2 = n_2 - 1$ are called the degrees of freedom of these estimates.

While defining the statistic F, the larger of the two variances is always placed in the numerator and the smaller in the denominator.

Senedecor has prepared tables that give, for different values of v1 and v2, the 5% and 1% critical values of F. If F denotes the observed (calculated) value and $F_{v1,v2}(\alpha)$ denotes the critical (tabulated value) of F at LOS, then $P\{F > F_{v1,v2}(\alpha)\} = \alpha$

NOTES

F test is not a two tailed test and is always a right tailed test , since F cannot be negative. Thus if $F > F_{v1,v2}(\alpha)$, then the difference between F and 1, i.e., the difference between S_1^2 and S_2^2 is significant at LOS 'α'. In other words the samples may not be regarded as dawn from the sample population with the same variance. If $F < F_{v1,v2}(\alpha)$, the difference is not significant at LOS α.

To test if two small samples have been drawn from the same normal population , it is not enough to test if their means differ significantly or not, because in this test we assumed that the two samples came from the same population or from populations with equal variance. So, before applying the t-test for the significance of the difference of two sample means, we should satisfy ourselves about the equality of the population variances by F-test

Example 1: A sample of size 13 gave an estimated population variance of 3.0, while another sample of size 15 gave an estimate of 2.5. Could both samples be from populations with same variance?

Solution:

$n_1 = 13$ $S_1^2 = 3.0$ and $v_1 = n_1 - 1 = 12$

$n_2 = 15$ $S_2^2 = 2.5$ and $v_2 = n_2 - 1 = 14$

$H_0 : \sigma_1^2 = \sigma_2^2$ i.e the two variances have been drawn from populations with the same variance.
$H_1 : \sigma_1^2 \neq \sigma_2^2$

Here $S_1^2 > S_2^2$

The -statistic is defined as a ratio of unbiased estimates of population variance.

Symbolically, $F = \frac{S_1^2}{S_2^2}$

$$F = \frac{3.0}{2.5} = 1.2$$

$v_1 = 12$ and $v_2 = 14$

$F(v_1 = 12 , v_2 = 14)$ at 5% LOS = 2.53 from the table value.

The calculated value of F is < the tabulated value
Therefore H_0 is accepted.

i.e., the two samples could have come from two normal populations with the same variance.

NOTES

Example 2: From the following data test if the difference between the variances is significant at 5% level of significance.

Sum of the squares of deviation from the mean	84.4	102.6
Size	8	10
Sample	A	B

Solution:

$H_0 : \sigma_1^2 = \sigma_2^2$ i.e the two variances have been drawn from populations with the same variance.

$H_1 : \sigma_1^2 \neq \sigma_2^2$

Given $\Sigma (x_1 - \overline{x}_1)^2 = 84.4$ and $\Sigma (x_2 - \overline{x}_2)^2 = 102.6$

$$S_1^2 = \frac{\Sigma (x_1 - \overline{x}_1)^2}{n_1 - 1} \quad \text{and} \quad S_2^2 = \frac{\Sigma (x_2 - \overline{x}_2)^2}{n_2 - 1}$$

$S_1^2 = 84.4 / 7 = 12.06$ and $S_2^2 = 102.6/ 9 = 11.4$

$S_1^2 > S_2^2$

The -statistic is defined as a ratio of unbiased estimates of population variance.

Symbolically, $F = \frac{S_1^2}{S_2^2}$

$$F = \frac{12.06}{11.4} = 1.058$$

$\nu_1 = n_1 - 1 = 7$ and $\nu_2 = n_2 - 1 = 9$

$F(\nu_1 = 7 , \nu_2 = 9)$ at 5% LOS = 3.29 from the table value.

The calculated value of F is < the tabulated value

Therefore H_0 is accepted.

i.e., the two samples could have come from two normal populations with the same variance.

Example 3: Time taken by workers in performing a job are given below: –

Method	1	20	16	26	27	23	22	
Method	2	27	33	42	35	32	34	38

Test whether there is any significant difference between the variances of time distribution.

NOTES

Solution:

Let us first calculate the variance of the samples.

Sample I x-22			Sample II y-34		
x	d	d^2	y	d	d^2
20	-2	4	27	-7	49
16	-6	36	33	-1	1
26	4	16	42	8	64
27	5	25	35	1	1
23	1	1	32	-2	4
22	0	0	34	0	0
		38	4	16	
134	2	82	241	3	135

$n_1 = 6 \qquad n_2 = 7$

$x = \Sigma x / n_1 = 134 / 6 = 22.33$

$y = \Sigma y / n_2 = 241 / 7 = 34.43$

Let $d_1 = x - 22$ $d_2 = y - 34$

$$s_1^2 = \frac{\Sigma d_1^2}{n_1} - \left(\frac{\Sigma d_1}{n_1}\right)^2$$

$s_1^2 = 82 / 6 - (2 / 6)^2 = 13.67 - 0.44 = 13.23$

$$s_2^2 = \frac{\Sigma d_2^2}{n_2} - \left(\frac{\Sigma d_2}{n_2}\right)^2$$

$s_2^2 = 135 / 7 - (3 / 7)^2 = 19.29 - 0.18 = 19.11$

$$S_1^2 = \frac{n_1 s_1^2}{n_1 - 1} = \frac{6 \times 13.23}{5} = 15.88 \quad \text{and} \quad S_2^2 = \frac{n_2 s_2^2}{n_2 - 1} = \frac{7 \times 19.11}{6} = 22.3$$

Here $S_2^2 > S_1^2$

$H_0 : \sigma_1^2 = \sigma_2^2$ i.e the two variances have been drawn from populations with the same variance.

$H_1 : \sigma_1^2 \neq \sigma_2^2$

NOTES

The -statistic is defined as a ratio of unbiased estimates of population variance.
Symbolically, $\mathbf{F = \frac{S_2^2}{S_1^2}}$

$$F = \frac{22.30}{15.88} = 1.40$$

$v_1 = n_1 - 1 = 6$ and $v_2 = n_2 - 1 = 5$

$F(v_1 = 6, v_2 = 5)$ at 5% LOS = 4.28 from the table value.
The calculated value of F is < the tabulated value
Therefore H_0 is accepted.
i.e., the two samples could have come from two normal populations with the same variance.

Example 4: Two random samples drawn from normal population are –

Sample I	20	16	26	27	23	22	18	24	25	19		
Sample II	27	33	42	35	32	34	38	28	41	43	30	37

Obtain estimates of variances of the population and test whether the two populations have the same variances.

Solution:
Let us first calculate the variance of the samples.

Sample I x-22			Sample II y-35		
x	d	d^2	y	d	d^2
20	-2	4	27	-8	64
16	-6	36	33	-2	4
26	4	16	42	7	49
27	5	25	35	0	0
23	1	1	32	-3	9
22	0	0	34	-1	1
18	-4	16	38	3	9
24	2	4	28	-7	49
25	3	9	41	6	36
19	-3	9	43	8	64
			30	-5	25
			37	2	4
220	0	120	420	0	314

NOTES

$n_1 = 10 \qquad n_2 = 12$

$\bar{x} = \Sigma x / n_1 = 220 / 10 = 22$

$\bar{y} = \Sigma y / n_2 = 420 / 12 = 35$

Let $d_1 = x - 22$ $d_2 = y - 35$

$$s_1^2 = \frac{\Sigma d_1^2}{n_1} - \left(\frac{\Sigma d_1}{n_1}\right)^2$$

$s_1^2 = 120 / 10 - 0 = 12$

$$s_2^2 = \frac{\Sigma d_2^2}{n_2} - \left(\frac{\Sigma d_2}{n_2}\right)^2$$

$s_2^2 = 314 / 12 - 0 = 26.17$

$S_1^2 = \frac{n_1 s_1^2}{n_1 - 1} = \frac{10 \times 12}{9} = 13.33$ and $S_2^2 = \frac{n_2 s_2^2}{n_2 - 1} = \frac{12 \times 26.17}{11} = 28.55$

Here $S_2^2 > S_1^2$

$H_0 : \sigma_1^2 = \sigma_2^2$ i.e the two variances have been drawn from populations with the same variance.

$H_1 : \sigma_1^2 \neq \sigma_2^2$

The -statistic is defined as a ratio of unbiased estimates of population variance.

Symbolically, $F = \frac{S_2^2}{S_1^2}$

$$F = \frac{28.55}{13.33} = 2.14$$

$v_1 = n_1 - 1 = 11$ and $v_2 = n_2 - 1 = 9$

$F(v_1 = 11, v_2 = 9)$ at 5% LOS = 3.10 from the table value.

The calculated value of F is < the tabulated value

Therefore H_0 is accepted.

i.e., the two samples could have come from two normal populations with the same variance.

NOTES

Example 5: Values of a variate in two samples are given below –

Sample I	5	6	8	1	12	4	3	9	6	10
Sample II	2	3	6	8	1	10	2	8		

Test the significance of the difference between the two sample means and the two sample variances.

Solution:

Let us first calculate the variance of the samples.

Sample I		Sample II	
x	x^2	y	y^2
5	25	2	4
6	36	3	9
8	64	6	36
1	1	8	64
12	144	1	1
4	16	10	100
3	9	2	4
9	81	8	64
6	36		64
10	100		
64	512	40	282

$n_1 = 10 \qquad n_2 = 8$

$\bar{x} = \Sigma x / n_1 = 64 / 10 = 6.4$

$\bar{y} = \Sigma y / n_2 = 40 / 8 = 5$

Let $d_1 = x - 22$ $d_2 = y - 35$

$$s_1^2 = \frac{\Sigma x_1^2}{n_1} - \left(\frac{\Sigma x_1}{n_1}\right)^2$$

$s_1^2 = 512 / 10 - (64 / 10)^2 = 51.2 - 40.96 = 10.24$

NOTES

$s_2^2 = \frac{\Sigma x_2^2}{n_2} - \frac{\Sigma x_2}{n_2}^2$

$s_2^2 = 282 / 8 - (40 / 8)^2 = 35.25 - 25 = 10.25$

case 1: Test for mean: The samples are small and so we apply t-test.

$H_0 : \mu_1 = \mu_2$ (The means of the two population are equal)

$H_1 : \mu_1 \neq \mu_2$ (The means of the two population are not equal)

The test statistic is given by $t = \frac{x_1 - x_2}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$.

where $S^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$

$n_1 + n_2 - 2$ is the number of degrees of freedom of the statistic.

$S^2 = \frac{10 \times 10.24 + 8 \times 10.25}{10 + 8 - 2} = 11.525$

$S = 3.395$

Therefore $t = \frac{6.45 - 5}{3.395\sqrt{(1/10 + 1/8)}} = 0.87$

$|t| = 0.87$

number of degrees of freedom, ndf = 10 + 8– 2 = 16

The table value of t for 16 df at 5% level = 2.12

The calculated value of t < the table value of t.

H_0 is accepted at 5% level. Hence there is no significant difference between the means of the population.

Case 2: Test for variance

$n_1 = 10 \qquad n_2 = 8$

$s_1^2 = 10.24 \quad s_2^2 = 10.25$

$S_1^2 = \frac{n_1 s_1^2}{n_1 - 1} = \frac{10 \times 10.24}{9} = 11.38$ and $S_2^2 = \frac{n_2 s_2^2}{n_2 - 1} = \frac{8 \times 10.25}{7} = 11.71$

NOTES

Here $S_2^2 > S_1^2$

$H_0 : \sigma_1^2 = \sigma_2^2$ i.e the two variances have been drawn from populations with the same variance.
$H_1 : \sigma_1^2 \neq \sigma_2^2$

The -statistic is defined as a ratio of unbiased estimates of population variance.
Symbolically, $F = \frac{S_2^2}{S_1^2}$

$$F = \frac{11.71}{11.38} = 1.03$$

$\nu_1 = n_1 - 1 = 7$ and $\nu_2 = n_2 - 1 = 9$

$F(\nu_1 = 11, \nu_2 = 9)$ at 5% LOS = 3.29 from the table value.

The calculated value of F is < the tabulated value
Therefore H_0 is accepted.

i.e., the two samples could have come from two normal populations with the same variance.

Example 6 : Two random samples gave the following data –

	Size	Mean	Variance
Sample I	8	9.6	1.2
Sample II	11	16.5	2.5

Can we conclude that two samples have been drawn from the same normal population.

Solution:

To conclude that the two samples have been drawn from the same population, we have to check first that the variances of the populations do not differ significantly and then check that the sample means (and hence the population means) do not differ significantly

$n_1 = 10 \qquad n_2 = 8$

$\overline{x}_1 = 9.6 \quad \overline{x}_2 = 16.5$

$s_1^2 = 1.2 \quad s_2^2 = 2.5$

NOTES

Case 1: Test for variance

$S_1^2 = \frac{n_1 s_1^2}{n_1 - 1} = \frac{8 \times 1.2}{7} = 1.37$ and $S_2^2 = \frac{n_2 s_2^2}{n_2 - 1} = \frac{11 \times 2.5}{10} = 2.75$

$H_0 : \sigma_1^2 = \sigma_2^2$ i.e the two variances have been drawn from populations with the same variance.

$H_1 : \sigma_1^2 \neq \sigma_2^2$

The -statistic is defined as a ratio of unbiased estimates of population variance.
Symbolically, $F = \frac{S_2^2}{S_1^2}$

$$F = \frac{2.75}{1.37} = 2.007$$

$v_1 = n_1 - 1 = 10$ and $v_2 = n_2 - 1 = 7$

$F(v_1 = 10, v_2 = 7)$ at 5% LOS = 3.64
from the table value.

The calculated value of F is < the tabulated value
Therefore H_0 is accepted.

i.e., the two samples could have come from two normal populations with the same variance.

Case 2: Test for mean: The samples are small and so we apply t-test.

$H_0 : \mu_1 = \mu_2$ (The means of the two population are equal)

$H_1 : \mu_1 \neq \mu_2$ (The means of the two population are not equal)
The test statistic is given by $t = \frac{x_1 - x_2}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

where $S^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$

$n_1 + n_2 - 2$ is the number of degrees of freedom of the statistic.

$S^2 = \frac{8 \times 1.2 + 11 \times 2.5}{17} = 11.24$

$S = 2.1823$

Therefore $t = \frac{9.6 - 16.5}{2.1823\sqrt{(1/8 + 1/11)}} = -10.05$

NOTES

|t| = 10.05
number of degrees of freedom, ndf = 8 + 11 – 2 = 17

The table value of t for 17 df at 5% level = 2.11

The calculated value of t > the table value of t.

H_0 is rejected at 5% level. Hence there is significant difference between the means of the population.

Hence the two samples could not have been drawn from the same normal population.

How you understood ?

1.State the important properties of the F-distribution.
2.What is the use of F-distribution?
3.Write down the probability density function of the F-distribution.

TRY YOURSELF !

1) In a sample of 8 observations, the sum of the squared deviations of items from the Mean was 94.5. In another sample of 10 observations, the value was found to be 101.7. Test whether the difference in the variances is significant at 5% level.

2) Two samples were drawn from two normal populations and their values are

A	66	67	75	76	82	84	88	90	92		
B	64	66	74	78	82	85	87	92	93	95	97

Test whether the two populations have the same variance at 5% level of significance.

3) In tests given to two groups of students drawn from two different populations, the marks obtained were as follows

Group A	18	20	36	50	49	36	34	49	41
Group B	29	28	26	35	30	44	46		

NOTES

3.6 CHI SQUARE TEST

Karl Pearson has shown that if $X_1, X_2, X_3, \ldots X_n$ are n independent normal variables with means $\mu_1,\mu_2,\ldots\mu n$ and standard deviation $\sigma_1,\sigma_2, \ldots . \sigma_n$ then the random variable defined by $\chi^2 = X_1^2 + X_2^2 + \ldots . + X_n^2$ has a probability distribution called χ^2 distribution with n degrees of freedom. Here n is only the number of independent variables under consideration. The importance of this distribution is that obeys additive property.

❖ *ADDITIVE PROPERTY*

If $\chi_1^2, \chi_2^2, \ldots . \chi_k^2$ are k independent χ^2 random variables with $n_1, n_2, \ldots, n_k$ degrees of freedom then their sum $\chi^2 = \chi_1^2 + \chi_2^2 + \ldots . + \chi_k^2$ is also a χ^2 random variable with
$n_1 + n_2 + \ldots + n_k$ number of degrees of freedom.

Pearson has shown that χ^2 – statistic is useful for comparison of observed frequencies with theoretical frequencies and to draw the decision whether there is any significant difference between these two sets. In this context χ^2 is called a non-parametric test.

❖ *Pearson's Statistics*

For testing the significance of difference between observed and expected frequencies under the null hypothesis that the difference is insignificant, Pearson has constructed the statistic that

$$\chi^2 = \Sigma \left[\frac{(O_i - E_i)^2}{E_i} \right]$$

Here O_i are the observed frequencies and E_i are the expected frequencies. The expected frequencies can be calculated on the assumption of H_0.

Pearson has shown that for large sample, this statistic follows χ^2 distribution with n-1 degrees of freedom. The sampling distribution of χ^2 is given by

$f(\chi^2) = c(\chi^2)^{n/2-1} \exp(-\chi^{n/2})$

where the constant c is to be determined such that $\int f(\chi^2) d\chi^2 = 1$

The χ^2 distribution has only one parameter *v* called the number of degrees of freedom. For each value of *v* χ^2 has different curve. For small values of *v* the curve is skewed the the right. For large values of *v*, χ^2 distribution is closed approximated to the normal distribution.

NOTES

❖ *USES OF χ^2 TEST*

The following are the uses of χ^2 statistic –

(1) It is used to test the goodness of fit of a distribution.
(2) It is used to test the significance of the difference between the observed frequencies in a sample and the expected frequencies obtained from the theoretical distribution.
(3) It is used to test the independence of the attributes.
(4) In the case of small samples (where the population standard deviation is not known), χ^2 statistic is used to test whether a specified value can be the population variance σ^2.

3.6.1 χ^2 Test for Goodness of Fit

Procedure for testing the significance of the difference between the observed and expected frequencies.

H_0 – There is no significant differences between the observed and the expected frequencies.

H_1 – There is significant difference between the observed and the expected frequencies.

The test statistic is

$$\chi^2 = \Sigma \frac{(O - E)^2}{E}$$

the expected frequencies are determined on the assumption that H_0 is true.

The number of degrees of freedom = n-1 where n is the number of classes. From the χ^2 table we can find for a given degrees of freedom the table value of χ^2 for a given significance level (say $\alpha = .05$ or $\alpha = 0.01$)

If the calculated value of χ^2 < the table value of χ^2 , H_0 is accepted at the significance level α.

If the calculated value of χ^2 > the table value of χ^2 , H_0 is rejected at the significance level α.

Note - For testing the goodness of fit of a distribution by assuming H_0 as some specific distribution, Binomial, Poisson etc, we calculate the theoretical frequencies and adopt the procedure given above to test whether the assumed distribution is a better fit for the observed frequencies.

NOTES

✓ Conditions for the validity of χ^2 test

1. The number of observations N in the sample must be reasonably large, say $\frac{24}{12}$ 50
2. Individual frequencies must not be too small, i.e., O $\frac{24}{12}$ 10. In case of O $\frac{c}{\lambda}$ 10, it is combined with the neighbuoring frequencies, so that the combined frequency is $\frac{24}{12}$ 10.
3. The number of classes n must be neither too small nor too large i.e., 4 $\frac{c}{\lambda}$ n $\frac{c}{\lambda}$ 16.

3.6.2 Test of independence of attributes

Another important application of the χ^2 distribution is the testing of independence of attributes (attributes are characters which are non measurable – for eg. Sex, Employment, Literacy etc are all attributes). Suppose we want to test whether sex and employment are associated. In this case take a random sample from the population an classify the sample as given in the following table. The numbers in the table denote the frequencies (number of persons possessing the attribute)

	Male	Female	Total
Employed	50	20	70
Unemployed	15	15	30
Total	65	35	100

This type of table which has one basis of classification across column and another across row is known as contingency table. The above table has 2 rows and 2 columns and hence is called as 2 X 2 contingency table. A table which as r rows and s columns is called a
r X s contingency table.

In testing the hypothesis the null hypothesis is taken as " employment is independent of sex" where as the alternate hypothesis is " employment is not independent of sex".

Then comes the question of determining the expected frequencies.

Assuming that H_0 is true, the totals are all kept the same.

For example, the expected frequency for the 1st cell in the above table, is determined by the formula :

$$\frac{\textbf{Row total} \times \textbf{Column total}}{\textbf{Grand Total}}$$

$$= \frac{70 \text{ X } 65}{100} = 45.5$$

NOTES

The other theoretical frequencies are determined on the same lines –

	Male	Female	Total
Employed	45.5	24.5	70
Unemployed	24.5	5.5	30
Total	65	35	100

It can be checked that by determining the only one cell frequency the other expected frequencies can be easily obtained from the column and row totals. Thus in a 2 X 2 contingency table the number of degrees of freedom is (2 – 1)X(2 – 1) = 1. In general in a r X s contingency table the number of degrees of freedom is (r – 1) X (s – 1).

Test procedure –

Step 1 – write down the null hypothesis.
Step 2 – write down the alternate hypothesis.
Step 3 – calculate the theoritial frequencies for the contingency table.

Step 4 – calculate $\chi^2 = \Sigma \frac{(O - E)^2}{E}$

Step 5 – write down the number of degrees of freedom.
Step 6 - draw the conclusion on the hypothesis by comparing the calculated values of χ^2 with the table value of χ^2.

Note the value of χ^2 statistic for a 2 X 2 contingency table can also be calculated using the formula given below –

	A	A	Total
B	a	b	a + b
B	c	d	c + d
Total	a + c	b + d	N

$$\chi^2 = \frac{N(ad - bc)^2}{(a + b)(c + d)(a + c)(a + d)}$$

3.6.3 Test for a specified population variance

Let $\{x_1, x_2, \ldots x_n\}$ be a random sample of size n drawn from a normal population. We want to test, based on the sample, whether the population variance can be σ_0^2. Let us now give the procedure for the test
$H_0: \sigma^2 = \sigma_0^2$.

NOTES

$H_1: \sigma^2 \neq \sigma_0^2$.
The test statistic is χ^2.

On the assumption that H_0 is true, it has been shown that the statistic $\chi^2 = ns^2 / \sigma^2$ has a
χ^2 distribution with (n - 1) degrees of freedom.
In this formula n is the sample size , s^2 is the variance, σ^2 is the population variance

We can determine the table value of χ^2 for (n - 1) degrees of freedom.
Accept H_0 if the calculated value of χ^2 < the table value. Reject H_0 if the calculated value of χ^2 > the table value.

Example 1: A company keeps records of accidents. During a recent safety review, a random sample of 60 accidents was selected and classified by the day of the week on which they occurred.

Day	Mon	Tue	Wed	Thu	Fri
No. of Accidents	8	12	9	14	17

Test whether there is any evidence that the accidents are more likely on some days than others.

Solution:

H_0 – Accidents are equally likely to occur on nay day of the week.

H_1 – Accidents are not equally likely to occur on the days of the week.

Total number of accidents = 60

On the assumption H_0, the expected number of accidents on any day

$= \frac{60}{5} = 12$

Let O denote the observed frequency and E denote the expected frequency

O	E‘	O - E	$(O - E)^2$
8	12	-4	16
12	12	0	0
9	12	-3	9
14	12	2	4
17	12	5	25
60	60		54

NOTES

$$\chi^2 = \Sigma \left[\frac{(O - E)^2}{E}\right]$$

$$= \frac{54}{12} = 4.5$$

N = number of classes = 5
Thus number of degrees of freedom = n – 1 = 5 – 1 = 4

For 4 degrees of freedom the table value of χ^2 is 9.4888.
But the calculated value of χ^2 is 4.5

Thus the calculated value of χ^2 < the table value of χ^2

Hence H_0 is accepted at 5% level. This means that accidents are equally likely to occur on any day of the week.

Example 2: A company produces a product of 4 sizes : small, medium, large and extra large. In the past the demand for these sizes has been fairly constant at 20% for small, 45% for medium, 25% for large and 10% for extra large. A random sample of 400 recent sales included 66 small, 172 medium, 109 large and 53 extra large. Test whether there is evidence of significant change in demand for the different sizes.

Solution:

H_0 – There is no evidence of significant change in demand for the different sizes.

H_1 – There is evidence of significant change in demand for the different sizes.

The expected frequencies are –

$\frac{20}{100} \times 400$, $\frac{45}{100} \times 400$, $\frac{25}{100} \times 400$ and $\frac{10}{100} \times 400$

i.e., 80,180,100,40

Size	O	E‘	O - E	(O - E)^2	$\left(\frac{O - E}{E}\right)^2$
Small	66	80	-14	196	2.45
Medium	172	180	-8	64	0.356
Large	109	100	9	81	0.810
Extra Large	53	40	13	169	4.225
	400				7.841

NOTES

$\chi^2 = \Sigma \left[\frac{(O - E)^2}{E}\right] = 7.841$

Number of degrees of freedom = 4 – 1 = 3

For 3 degrees of freedom the table value of χ^2 at 5% level is 7.81.
But the calculated value of χ^2 is 7.841

Thus the calculated value of χ^2 > the table value of χ^2

Hence H_0 is rejected at 5% level. This means that there is evidence of significant change in demand for the different sizes.

Example 3: In 20 throws of a single die the following distributions of faces was observed

Face	1	2	3	4	5	6
Frequency	30	25	18	10	22	15

Can you say that the die is unbiased?

Solution:

H_0 – The die is unbiased.

H_1 – The die is biased.

On the assumption H_0, the expected frequency for each face = $120 \times \frac{1}{6} = 20$

Face	O	E	O - E	(O - E)^2
1	30	20	10	100
2	25	20	5	25
3	18	20	-2	4
4	10	20	-10	100
5	22	20	2	4
6	15	20	-5	25
		120		258

If E is same for all u need not have a separate column to find $\left(\frac{O - E}{E}\right)^2$

$\chi^2 = \Sigma \left[\frac{(O - E)^2}{E}\right] = \frac{258}{20} = 12.9$

NOTES

Number of degrees of freedom = n – 1 = 6 – 1 = 5

For 5 degrees of freedom the table value of χ^2 at 5% level is 11.07
But the calculated value of χ^2 is 12.9

Thus the calculated value of χ^2 > the table value of χ^2

Hence H_0 is rejected at 5% level. Hence the die can be regarded as biased.

Example 4: A sample analysis of examination results of 500 students was made. It was found that 220 students have failed, 170have secured a third class, 90 have secured a second class, and the rest, a first class. Do these figures support the general brief that the above categories are in the ratio 4:3:2:1 respectively?

Solution:

H_0 – The results in the four categories are in the ratio 4:3:2:1.
H_1 – The results in the four categories are not in the ratio 4:3:2:1.

On the assumption $H_{0,t}$ he expected frequencies are –

$\frac{40}{10}$ x 500, $\frac{3}{10}$ x 500, $\frac{2}{10}$ x 500 and $\frac{1}{10}$ x 500

i.e., 200, 150, 100, 50.

	O	E	O - E	(O - E)^2	$\frac{(O - E)^2}{E}$
Failures	220	200	20	400	2.000
III class	170	150	20	400	2.667
II class	90	100	-10	100	1.000
I class	20	50	-30	900	18.000
	500				23.667

$\chi^2 = \Sigma \left[\frac{(O - E)^2}{E}\right] = 23.667$

Number of degrees of freedom = 4 – 1 = 3

For 3 degrees of freedom the table value of χ^2 at 5% level is 7.81.
But the calculated value of χ^2 is 23.667

Thus the calculated value of χ^2 > the table value of χ^2

NOTES

Hence H_0 is rejected at 5% level. This means the results in the four categories are not in the ratio 4:3:2:1.

Example 5: The following table shows the distribution of digits in numbers chosen at random from a telephone directory

Digit	0	1	2	3	4	5	6	7	8	9
Frequency	1026	1107	997	966	1075	933	1107	972	964	853

Test whether the digits may be taken t occur equally frequently in the directory.

Solution:

H_0 – The digits occur equally frequently in the directory.

H_1 – The digits do not occur equally frequently

On the assumption H_0, the expected frequency for each face = $\frac{10000}{10} = 1000$

Digit	O	E	O - E	$(O - E)^2$
0	1026	1000	26	0.676
1	1107	1000	107	11.449
2	997	1000	3	0.009
3	966	1000	34	1.156
4	1075	1000	75	5.625
5	933	1000	67	4.489
6	1107	1000	107	11.449
7	972	1000	28	0.784
8	964	1000	36	1.296
9	853	1000	147	21.609
	10000	120		58.542

If E is same for all u need not have a separate column to find $\frac{(O - E)^2}{E}$

$$\chi^2 = \Sigma \left[\frac{(O - E)^2}{E}\right] = 58.542$$

Number of degrees of freedom = n – 1 = 10 – 1 = 5

For 5 degrees of freedom the table value of χ^2 at 5% level is 16.919

But the calculated value of χ^2 is 58.542.

NOTES

Thus the calculated value of χ^2 > the table value of χ^2

Hence H_0 is rejected at 5% level. The digits are not uniformly distributed in the directory.

Example 6: A set of 5 identical coins is tossed 320 times and the number of heads appearing each time is recorded.

0	1	2	3	4	5
14	45	80	772	61	8

Test whether the coins are unbiased at 5% level of significance.

Solution:

H_0: coins are unbiased (P(getting head) = p = ½ , q = 1/6)
H_1 coins are not biased

On the assumption H_0, the probability of getting exactly 'r' successes = $5Crp^rq^{5-r}$ (r = 0, 1, 2, …5)
Therefore the expected number of times in which exactly 'r' successes are obtained

= 320 $\times 5C_r\, p^rq^{5-r}$
= 10, 50, 100, 100, 5, 10

No: of heads	O	E	O - E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
0	14	10	4	16	1.60
1	45	50	-5	25	0.50
2	80	100	-20	400	4.0
3	112	100	12	144	1.44
4	1	5	11	121	2.42
5	8	10	-2	4	0.40
	320	320			10.36

$\chi^2 = \sum\left[\frac{(O-E)^2}{E}\right] = 10.36$

Number of degrees of freedom = n – 1 = 6 – 1 = 5
Table value of χ^2 for 5 at 5% level = 11.07

Since the calculated value of χ^2 is less than the table value of χ^2, H_0 is accepted at 5% level.

NOTES

Hence the coins are unbiased.

Example 7: A survey of 320 families with five children each revealed the following distribution.

No: of boys	0	1	2	3	4	5
No: of girls	5	4	3	2	1	0
No: of families	12	40	88	110	56	14

Is the result consistent with the hypothesis that male and female births are equally probable?

Solution:

H_0: male and female births are equally probable (P(male birth) = p = ½ , q = 1/6)
H_1: male and female births are not equally probable

On the assumption H_0, the probability that a family of 5 children has r male children = $5Crp^rq^{5-r}$ (r = 0, 1, 2, …5)
Therefore the expected number of times in which exactly 'r' successes are obtained
= 320 $\times 5C_r\, p^rq^{5-r}$
= 10, 50, 100, 100, 5, 10

No: of males	O	E	O - E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
0	12	10	2	4	0.4
1	40	50	-10	100	0.5
2	88	100	-22	484	4.84
3	110	100	10	100	1
4	56	50	6	36	0.72
5	14	10	4	16	1.6
	320	320			7.16

$\chi^2 = \sum\left[\frac{(O-E)^2}{E}\right] = 7.16$
Number of degrees of freedom = n – 1 = 6 – 1 = 5
Table value of χ^2 for 5 at 5% level = 11.07

Since the calculated value of χ^2 is less than the table value of χ^2, H_0 is accepted at 5% level.

Hence the male and female births are equally probable.

NOTES

Example 8: Fit a binomial distribution for the following data and also test the goodness of fit

x	0	1	2	3	4	5	6	Total
f	5	18	28	12	7	6	4	80

Solution:

H_0: the given distribution is approximately a binomial distribution

To find the binomial distribution $N(q + p)^n$, which fits the given data, we require p.

We know that the mean of the binomial distribution is np, from which we can find p. Now the mean of the given distribution is found out and is equated to np.

x	0	1	2	3	4	5	6	Total
f	5	18	28	12	7	6	4	80
fx	0	18	56	36	28	30	24	192

$\bar{x} = \frac{\Sigma fx}{\Sigma f} = 192/80 = 2.4$

i.e., np = 2.4 or 6p = 2.4, since the maximum value taken by x is n.
p = 0.4 and hence q = 0.6

The expected frequencies are given by
$= 80 \times {}^6C_r\, p^r q^{6-r}$ (r = 0, 1, 2, 3, 6) = 3.73 ≅ 4, 14.93 ≅ 15, 24.88 ≅ 25, 22.12 ≅ 22, 11.06 ≅ 11, 2.95 ≅ 3, 0.33 ≅ 0

O	5	18	28	12	7	6	4
E	4	15	25	22	11	3	0

The first class is combined with the second and the last two classes are combined with the last but second class in order to make the expected frequency in each class greater than or equal to 10. Thus after grouping

O	23	28	12	17
E	19	25	22	14

NOTES

O	E	O - E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
23	19	4	16	0.8421
28	25	3	9	0.36
12	22	10	100	0.5455
17	14	3	9	0.6429
				6.39

$$\chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right] = 6.39$$

Number of degrees of freedom = n – k = 4 – 2 = 2
Table value of χ^2 for 2at 5% level = 5.99

Since the calculated value of χ^2 is > the table value of χ^2, H_0 is rejected at 5% level.
i.e., the binomial fit for the given distribution is not satisfactory.

Example 9: Fit a Poisson distribution for the following data and also test the goodness of fit

x	0	1	2	3	4	5	Total
f	142	156	69	27	5	1	400

Solution:

H_0: the given distribution is approximately a Poisson distribution

To find the Poisson distribution whose probability law is

$$P(X = r) = \frac{e^{-\lambda}\lambda^r}{r!}, \quad r = 0, 1, 2, \ldots$$

We require λ the mean of the Poisson distribution.
We will find the mean of the given data and assume it as λ.

x	0	1	2	3	4	5	Total
f	142	156	69	27	5	1	400
fx	0	156	138	81	20	5	400

$$\bar{x} = \frac{\Sigma fx}{\Sigma f} = 400/400 = 1$$

NOTES

The expected frequencies are given by

$\frac{Ne^{-\lambda}\lambda^r}{r!}$, r = 0, 1, 2, . …= > $\frac{400e^{-\lambda}\lambda^r}{r!}$, r = 0, 1, 2, .

= 147.15 ≅ 147, 147.15 ≅ 147, 73.58 ≅ 74 , 24.53 ≅ 25, 6.13 ≅ 6, 1.23 ≅ 1

O	142	156	69	27	5	1
E	147	147	74	25	6	1

The last three classes are combined into one, so that the expected frequency in the class may be greater than 10. Thus after regrouping, we have

O	142	156	69	33
E	147	147	74	32

O	E	O - E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
142	147	-5	25	0.17
156	147	9	81	0.551
69	74	-5	25	0.027
33	32	1	1	0.0312
				1.09

$\chi^2 = \sum\left[\frac{(O-E)^2}{E}\right] = 1.09$

Number of degrees of freedom = n – k = 4 – 2 = 2
Table value of χ^2 for 2at 5% level = 5.99

Since the calculated value of χ^2 is < the table value of χ^2, H_0 is accepted at 5% level. i.e., the Poisson fit for the given distribution is satisfactory.

Problems on Independence of attributes

Example 10 : A random sample of employees of a large company was selected and the employees were asked to complete a questionnaire. One question asked was whether the employee was in favour of the introduction of flexible working hours. The following table classifies the employees by their response and by their area of work.

NOTES

Response	Area of work	
	Production	Non Production
In favour	129	171
Not in favour	1	69

Test whether there is evidence of a significant association between the response and the area of work?

Solution:

H_0: There is no evidence of a significant association between the response and the area of work

H_1: There is an evidence of a significant association between the response and the area of work

Now we have to calculate the expected frequencies to apply the χ^2 test.

On the assumption of H_0, the expected frequency for the class ''production an in favour" is given by

$$\frac{(A) \times (B)}{N} = \frac{160 \times 300}{400} = 120$$

Similarly we can calculate the other expected frequencies .

The other expected frequencies are

$$\frac{240 \times 300}{400} = 180, \quad \frac{160 \times 100}{400} = 40, \quad \frac{240 \times 100}{400} = 60$$

Table showing observed frequencies

Response	Production	Non Production	Total
In favour	129	171	300
Not in favour	1	69	100
Total	160	240	400

Table showing Expected frequencies

NOTES

Response	Production	Non Production	Total
In favour	129	180	300
Not in favour	40	60	100
Total	160	240	400

O	E	O - E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
129	120	9	81	0.675
171	180	-9	81	.450
31	40	-9	81	2.025
69	60	9	81	1.350
400	400			4.500

$$\chi^2 = \sum \left[\frac{(O-E)^2}{E}\right] = 4.5$$

Number of degrees of freedom = (2 – 1)(2 - 1) = 1
Table value of χ^2 for 1 at 5% level = 3.81

Since the calculated value of χ^2 is greater than the table value of χ^2, H_0 is rejected at 5% level.

Hence there is evidence for a significant association between response and the area of work

Example 11 : Can vaccination be regarded as a preventive measure of small-pox evidenced by the following data? ''Of 1482 persons exposed to small-pox in a locality 368 in all were attacked". Given the chi-square value at 5% level of significance fro 1 df is 3.84

Solution:

H_0: There is no evidence that vaccination can be regarded as a preventive measure of small-pox
H_1: There is evidence that vaccination can be regarded as a preventive measure of small-pox

NOTES

Table showing observed frequencies

	Vaccinated	Non Vaccinated	Total
Attacked	35	333	368
Not Attacked	308	806	1114
Total	343	1139	1482

Table showing Expected frequencies

Response	Vaccinated	Non Vaccinated	Total
Attacked	$\frac{343 \times 368}{1482} = 85$	283	368
Not Attacked	258	856	1114
Total	343	1139	400

$\chi^2 = \sum\left[\frac{(O - E)^2}{E}\right] = 50.85$

Number of degrees of freedom = (2 – 1)(2 - 1) = 1
Table value of χ^2 for 1 at 5% level = 3.81

Since the calculated value of χ^2 is greater than the table value of χ^2, H_0 is rejected at 5% level.

Hence there is evidence for regarding vaccination as a preventive measure for small-pox.

Example 12: To test the efficiency of a new drug a controlled experiment was conducted wherein 300 patients were administered the new drug and 200 other patients were not given the drug. The patients were monitored and the results were obtained as follows:

	Cured	Condition worsened	No effect
Given the drug	200	40	60
Not given the drug	120	30	50

Use χ^2 test for finding the effect of the drug.

NOTES

Solution:

H_0: The drug is not effective

H_1: The drug is effective

Table showing observed frequencies

	Cured	Condition worsened	No effect	Total
Given the drug	200	40	60	300
Not given the drug	120	30	50	200
	320	70	110	500

Table showing Expected frequencies

	Cured	Condition worsened	No effect	Total
Given the drug	$\frac{320 \times 300}{500} = 192$	42	66	300
Not given the drug	128	28	44	200
	320	70	110	500

O	E	O - E	$(O - E)^2$	$\frac{(O - E)^2}{E}$
200	192	8	64	0.3313
40	42	-2	4	0.0952
60	66	-6	36	0.5454
120	128	-8	64	0.5000
30	28	2	4	0.1429
50	44	6	36	0.8182
500	500			2.4330

$$\chi^2 = \sum \left[\frac{(O - E)^2}{E}\right] = 2.43$$

NOTES

Number of degrees of freedom = (2 – 1)(3 - 1) = 2
Table value of χ^2 for 2 at 5% level = 5.991

Since the calculated value of χ^2 is < the table value of χ^2, H_0 is accepted at 5% level. Hence the drug is not effective.

Example 13 : A sample of hotels in a particular country was selected. The following table shows the number of hotels in each region of t h country and in each of four grades

Grade	Region		
	Eastern	central	Western
1 star	29	22	29
2 star	67	38	55
3 star	53	32	35
4 star	11	8.	21

Show that there is evidence of a significant association between region and grade of hotel in this country.

Solution:

H_0: There is no evidence for significant association between region and grade of hotel
H_1: There is evidence for significant association between region and grade of hotel

Table showing observed frequencies

	Region			Total
	Eastern	central	Western	
1 star	29	22	29	80
2 star	67	38	55	160
3 star	53	32	35	120
4 star	11	8.	21	40
	160	100	140	400

Table showing Expected frequencies

	Region			Total
	Eastern	central	Western	
1 star	32	0	28	80
2 star	64	60	56	160
3 star	48	30	42	120
4 star	16	10	14	40
	160	100	140	400

NOTES

O	E	O - E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
29	32	-3	9	0.281
22	20	2	4	0.200
29	28	1	1	0.036
67	64	3	9	0.141
38	40	-2	4	0.200
5	56	-1	1	0.018
53	48	5	25	0.521
32	30	2	4	0.133
35	42	-7	49	1.167
11	16	-5	25	1.562
8	10	-2	4	0.40
1	14	7	49	3.500
400	400			8.519

$$\chi^2 = \sum\left[\frac{(O-E)^2}{E}\right] = 8.159$$

Number of degrees of freedom = (4 – 1)(3 - 1) = 6
Table value of χ^2 for 6 at 5% level = 12.59

Since the calculated value of χ^2 is > the table value of χ^2, H_0 is rejected at 5% level. Hence there is evidence for significant association between region and grade of hotel.

Example 14 : A credit rating agency conducted a survey of customers and analyses them by occupation and credit risk. The results were as follows:

Credit rating	Administrative & clerical	Skilled manual	Semi-skilled & unskilled
High	60	50	10
Average	30	20	10
Poor	10	10	40

Test whether there is any association between occupation and credit rating?

Solution:

H_0: There is no association between occupation and credit rating
H_1: There is association between occupation and credit rating

NOTES

Table showing observed frequencies

Credit rating	Administrative & clerical	Skilled manual	Semi-skilled & unskilled	Total
High	60	50	10	120
Average	30	20	10	60
Poor	10	10	40	60
	100	80	60	240

Table showing Expected frequencies

Credit rating	Administrative & clerical	Skilled manual	Semi-skilled & unskilled	Total
High	$\frac{100 \times 120}{240} = 50$	40	30	120
Average	25	20	15	60
Poor	25	20	15	60
	100	80	60	240

O	E	O - E	$(O - E)^2$	$\frac{(O - E)^2}{E}$
60	50	10	100	2.00
50	40	10	100	2.50
10	30	-20	400	13.33
30	25	5	25	1.00
20	20	0	0	0.00
10	15	-5	25	1.67
10	25	-15	25	9.00
10	20	-10	100	5.00
40	15	25	625	41.67
400	400			76.17

$$\chi^2 = \sum \left[\frac{(O - E)^2}{E}\right] = 76.17$$

Number of degrees of freedom = (3 – 1)(3 - 1) = 4
Table value of χ^2 for 4 at 5% level = 9.49

NOTES

Since the calculated value of χ^2 is > the table value of χ^2, H_0 is rejected at 5% level. Therefore there is association between occupation and credit rating

.

Problems on test of specified population variance

Example 15: Weights in Kg of 10 students are given below:

38,40,45, 53, 47, 43, 55, 48, 52, 49

Can we say that the variance of the distribution of weights of all students from which the above sample of 10 students was drawn is equal to20 square kg?

Solution:

Here we have to apply the χ^2 –test for testing the significance of the difference between the sample variance and the population variance.

$H_0 : \sigma^2 = 20$ (there is no significant difference between the sample variance and the population variance)

$H_1 : \sigma^2 \neq 20$ (there is significant difference between the sample variance and the population variance)

x	d	d^2
38	-9	81
40	-7	49
45	-2	4
53	6	36
47	0	0
43	-4	16
55	8	4
48	1	1
52	5	25
49	2	4
470		280

$d = x - 47$

$s^2 = \frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2 = \frac{280}{10} - 0 = 28 \text{ kg}^2$

$\sigma^2 = 20 \text{ kg}^2$

$n = 10$

$\chi^2 = \frac{ns^2}{\sigma^2} = \frac{10 \times 28}{20} = 14$

NOTES

Number of degrees of freedom = 10 – 1= 9

Table value of χ^2 for 9 df at 5% level = 16.919
H_0 is accepted since the calculated value of χ^2 < the table value of χ^2.
Hence the population variance can be regarded as 20 square kg..

Example 16: A random sample of size 20 from a normal population gives a sample mean of 42 and sample SD of 6. test the hypothesis that the population SD is 9. Clearly state the alternative hypothesis you allow for and the level of significance.

Solution:

$H_0 : \sigma = 9$
$H_1 : \sigma \neq 9$

$s^2 = 6$, $\sigma^2 = 81$, $n = 20$

$$\chi^2 = \frac{ns^2}{\sigma^2} = \frac{20 \times 36}{81} = 8.89$$

Number of degrees of freedom = 20 – 1= 19

Table value of χ^2 for 19 df at 5% level = 30.144

H_0 is accepted since the calculated value of χ^2 < the table value of χ^2.
Therefore the population standard deviation can b regarded as 9.

How you understood ?

1.Define Chi-square distribution?
2.State the important properties of χ^2-distribution.
3.Give two uses of χ^2 distribution.
4.What is χ^2- test of goodness of fit?
5.What is contingency table.

TRY YOURSELF !

1) The theory predicts that the proportion of beans in 4 given groups should be 9:3:3:1. In an examination with 1600 beans, the number in the 4 groups were 882, 313, 287 and 118. Does the experimental result support the theory?

NOTES

2) 4 coins were tossed at a time and this operation is repeated 160 times. It is found that 4 heads occur 6 times, 3 heads occur 43 times, 2 heads occur 69 times and one head occurs 34 times. Discuss whether the coins be regarded as unbiased?

3) Five coins are tossed 256 times. The number of heads observed is given below. Examine if the coins are unbiased, by employing chi-square goodness of fit

No: of heads	0	1	2	3	4	5
Frequency	5	35	75	84	45	12

4) 2 groups of 100 people each were taken for testing the use of vaccine. 15 persons contracted the disease out of the inoculated persons, while 25 contracted the disease in the other group. Test the efficiency of the vaccine .

5) An insurance company advertises in the press a special pension plan for self-employed persons. The advertisement includes a coupon which enables interested persons to complete and return to the company. The company then posts to the enquiries to the initial information about the pension plan. If there is no response from the enquiries to the initial information, a second information pack is sent to the enquiries. Enquiries are divided by the company into three categories: definitely takes on plan, shows interests in plan, not interested. The company analysed a sample of 200 respondents to the initial advertisement i.e., those who returned the coupon. The following data was obtained.

	Responds to I mailing	Responds to II mailing	Telephone call made
Takes out plan	36	24	30
Shows interest	18	16	16
Not interested	6	20	34

Test whether there is any association between response and interest in the pension plan?

6) The heights of 10 randomly chosen college students in cm are 170, 165, 172, 168, 172, 164, 169, 167, 161, 163 Can we take the variance of heights of college students as 17 cm?

NOTES

REFERENCES:

1. T.Veerarajan, “Probability, statistics and Random Process “, Tata McGraw Hill, 2002.
2. P.Kandasamy, K. Thilagavathi and K. Gunavathi,”Probability, Random Variables and Random processors”, S. Chand, 2003.
3. S.C Gupta and V.K Kapoor,”Fundementals of Mathemetical Statistics”,Sultan Chand & Sons, 2002

NOTES

UNIT 4

RANDOM PROCESSES

- Introduction
- Random process
- Classification
- Stationary process
- Markov process
- Markov chain
- Poisson process

4. 1 INTRODUCTION

In unit I you have studied about random variables. Random variable is defined as a function of the sample points in a simple space. It does not include the concept of time. But in the real world, we come across many time varying functions which are quite are random. By extending the concept of a random variable to include time it is possible to define a random process. In the case of random variables, only a real number is assigned to each sample point of the sample space of a random experiment. So we denoted it as X(s), a real function of s alone. But in case of random process, a function of time is assigned to each sample point, based on some rule.So it is denoted as X(s,t).

In the simplest possible case a stochastic process amounts to a sequence of random variables known as a time series

4.2 LEARNING OBJECTIVES

The students will acquire

- Knowledge of random process concepts.
- Skills in handling situations involving random variable when a function of time is assigned .

NOTES

- Knowledge in making scientific judgments in the time of uncertainty and variation.

4.3 RANDOM PROCESS

4. 3.1 Definition:

A random process is a collection (or *ensemble*) of random variables {X(s,t)} that are functions of a real variable namely time t where s ε S (sample space) and t ε T (parameter set or index set)

Another definition for random process: A random process or stochastic process is defined as a family of random variables {X(t)} defined at different instants of time.

The set of possible values of any individual member of the random process is called state space. Any individual member itself is called a sample function or ensemble member or a realization of the process.

i. If s and t are fixed {X(s,t)} is a number.

ii. If t is fixed {X(s,t)} is a random variable.

iii. If s is fixed, {X(s,t)} is a single time function.

iv. If s and t are variables, {X(s,t)} is a collection of random variables that are time function.

Hereafter if the parameter set T is discrete, the random process will be noted by {X(n)} or $\{X_n\}$.

If the parameter set T is continuous, the process will be denoted by {X(t)}

4.4 CLASSIFICATION

Depending on the continuous or discrete nature of the state space and the parameter set T, a random process can be classified into four types:

i) If both T and S are discrete, the random process is called a discrete random sequence.

For example, if X_n represents the outcome of the nth toss of a fair die, then $\{X_n, n \overset{24}{12} 1\}$ is a discrete random sequence, since T = {1,2,3, . . .} and S = {1,2,3,4,5,6}.

ii) If T is discrete and S is continuous, the random process is called a continuous random sequence.

For example, If X_n represents the temperature at the end of the nth hour of a day, then $\{X_n, 1 \overset{c}{\lambda} n \overset{c}{\lambda} 24\}$ is a continuous random sequence, since temperature can take any value in an interval and hence continuous.

NOTES

iii) If T is continuous and S is discrete, the random process is called discrete random process.

For example, if X(t) represents the number of telephone calls received in the interval (0, t) then {X(t)} is a discrete random process, since S = {0,1,2,3,…..}.

iv) If both T and S are continuous, the random process is called continuous random process.

For example if x(t) represents the maximum temperature at a place in the interval (0, t), {X(t)} is a continuous random process.

The word discrete or continuous is used to refer the nature of S and the word sequence or process to refer the nature of T.

A random process is called a deterministic process if all the future values can be predicted from past observations. A random process is called a non-deterministic process if all the future values of any values of any sample function cannot be predicted from past observations.

Probability Distribution and Density functions

To each random variable we can define the probability distribution function $F_X(x_{1:}t_1)$ as $F_X(x_{1:}t_1) = P[X(t_1) \overset{c}{\lambda} x_1]$ for any real number x_1. This is called the first-order distribution function of random variable $X(t_1)$.

The first order probability density function $f_X(x_1, t_1)$ is defined as the derivative of the first order probability density function.

$$f_X(x_1, t_1) = \frac{d}{dx_1} F_X(x_{1:}t_1)$$

Joint Distribution

For two random variables $X(t_1)$ and $X(t_2)$ defined at two time instants t_1 and t_2 from the random process X(t) we can define the second order joint probability distribution function as

$$F_X(x_1,x_2:t_1,t_2) = P\{X(t_1) \overset{c}{\lambda} x_1, X(t_2) \overset{c}{\lambda} x_2\}$$

and the second order joint probability density function as

$$f_X(x_1,x_2: t_{1,}t_2) = \frac{\partial^2 F_X(x_1,x_2:t_1,t_2)}{\partial x_1 \partial x_2}$$

We can extend this to n random variables. So n^{th} order joint probability distribution is defined as

$$F_X(x_1,x_{2,}\ x_n:t_1,t_{2,\ldots.}t_n) = P\{X(t_1) \overset{c}{\lambda} x_1, X(t_2) \overset{c}{\lambda} x_2,.... X(t_n) \overset{c}{\lambda} x_n\}$$

And the n^{th} order joint probability density function as

NOTES

$$f_X(x_1,x_{2,}\ x_n:t_1,t_{2,\ldots..}t_n) = \frac{\partial^n \quad F_X(x_1,x_{2,}\ x_n:t_1,t_{2,\ldots..}t_n)}{\partial x_1 \partial x_2 \ldots\ldots \partial x_n}$$

Average values of Random process

Mean of the process {X(t)} is the expected value of a typical member X(t) of t he process

$$\text{i.e } \mu = E(X(t)),$$

Autocorrelation of the process {X(t)}, denoted by $R_{xx}(t_1,t_2)$ or $R_x(t_1,t_2)$ or $R(t_1,t_2)$, is the expected value of the product of any two members $X(t_1)$ and $X(t_2)$ of the process.

$$\text{i.e., } R(t_1,t_2) = E[X(t_1) \times X(t_2)]$$

Autocovaraince of the process {X(t)}, denoted by $C_{xx}(t_1, t_2)$ or $C_x(t_1,t_2)$ or $C(t_1,t_2)$ is defined as

$$C(t_1,t_2) = R(t_1,t_2) - \mu(t_1) \times \mu(t_2)$$

Correlation coefficient of the process {X(t)}, denoted by $\rho_{xx}(t_1,t_2)$ or $\rho_x(t_1,t_2)$ or $\rho(t_1,t_2)$, is defined as

$$\rho(t_1,t_2) = \frac{C(t_1,t_2)}{\sqrt{C(t_1,t_1)}\ \sqrt{C(t_2,t_2)}} \quad .$$

Where $C(t_1,t_2)$ is the variance of $X(t_1)$.

4.5 STATIONARITY

4.5.1 Stationary process:

If certain probability distribution or averages do not depend on time (t), then the random process {X(t)} is called stationary.

4.5.1.1 Stationary to order one

A random process is said to be stationary to order one if the first order density functions defined for all the random variables of the process are same.

In other words first order density function of the process, should not change with a shift in time origin.

$f_X(x_1: t_1) = f_X(x_{1,}\ t_1 + \delta)$ should be true for any t_1 and a time-shift δ.

As $f_X(x_1: t_1)$ is independent of t_1, obviously, the mean of each random variable is same. Hence the mean of the process is a constant.

$E[X(t)] = X = \text{constant}$

NOTES

4.5.1.2 Stationary to order two

A random process is said to be stationary to order two if for all t_1, t_2 and δ, its second order density function satisfy the condition

$$f_X(x_1, x_2: t_1,t_2) = f_X(x_1, x_2: t_1 + \delta, t_2 + \delta)$$

It is clear that, for a second- order stationary process , the second order joint probability function is a function of only time difference $t_2 - t_1$ and not on the absolute time. As it is possible to define first-order density functions of a second –process, is is clear that a second-order stationary process will also be a first order stationary process.

For a second order stationary process the two moments $E(X_1^2)$ and $E(X_2^2)$ do not change with time and are constants whereas the second order moment $E(X_1X_2) = E(X(t_1)X(t_2)) = Rxx(t_1,t_2)$ which is also called as auto correlation function of the process is a function of time difference at which the random variable X_1 and X_2 are defined.

$$E[X_1X_2] = E[X(t_1)X(t_2)] = R_{XX}(t_1,t_2)$$

If $t_2 - t_1 = \tau$, then $R_{XX}(t_1,t_2) = R_{XX}(\tau)$

4.5.2 Strongly stationary process (SSS PROCESS)

A random process is called a strongly stationary process or strict sense stationary process, if all its finite dimensional distributions are invariant under translation of time parameter.

That is if the joint distribution of $X(t_1)$, $X(t_2)$,…… $X(t_n)$ is the same as that of $X(t_1+h)$, $X(t_2+h)$,…… $X(t_n+h)$ for all $t_1, t_2, \ldots t_n$ and $h > 0$ and for all n=1, then the random process {X(t)}is called a SSS process. If the definition given above holds good for n = 1,2,…..k only and not for n>k, then the process is called kth order stationary.

Two real valued random process {X(t)}and {Y(t)}are said to be jointly stationary in the strict sense, if the joint distribution of X(t) and Y(t) are invariant under translation of time.

4.5.3 Wide Sense Stationary Process (WSS PROCESS)

A random process X(t) with finite first and second order moments is called weakly stationary process or covariance stationary process or wide- sense stationary process if its mean is a constant and auto correlation depends only on the time difference.

i.e., if $E\{X(t)\} = \mu$ and

$E\{ X(t) \times X(t - \tau) \} = R(\tau)$

NOTES

Note:

From the definitions given above, is clear that a SSS process with finite first order and second order moments is a WSS process, while a WSS process need not be a SSS process.

Two real valued random process {X(t)}and {Y(t)}are said to be jointly stationary in the wide sense, if each process is individually a WSS process, $R_{xy}(t_1,t_2)$ is a function of $(t_1 - t_2)$ only.

Evolutionary process

A random process that is not stationary in any sense is called an evolutionary process.

Example 1: The process {X(t)}whose probability function is given by

$$P\{X(t) = n\} = \frac{(at)^{n-1}}{(1+at)^{n+1}}, n = 1,2,3,\ldots.$$

$$= \frac{at}{(1+at)}, n = 0$$

show that it is not stationary

Solution:

The probability distribution if X(t) is

X(t) = n	0	1	2	3	…..
P_n:	at/(1 +at)	$1/(1 + at)^2$	$at/(1 + at)^3$	$(at)^2/(1+ at)^3$	

$$E(X(t)) = \sum_{n=0}^{\infty} np_n$$

$$= 1/(1 + at)^2 + 2at/(1 + at)^3 + 3(at)^2/(1 + at)^4 + \ldots\ldots$$

$$= 1/(1 + at)^2 [1 + 2\alpha + 3\alpha^2 + \ldots..], \text{ where } \alpha = at/(1 + at)$$

$$= 1/(1 + at)^2 [1 - \alpha]^{-2}$$

$$= [1/(1 + at)^2] (1 + at)^2 = 1$$

$$E(X^2(t)) = \sum_{n=0}^{\infty} n^2 p_n = \sum_{n=1}^{\infty} n^2 \frac{(at)^{n-1}}{(1+at)^{n+1}}$$

$$= 1/(1 + at)^2 \sum_{n=1}^{\infty} n(n + 1)[at/1+at]^{n-1} - \sum_{n=1}^{\infty} n[at/1+at]^{n-1}$$

$$= 1/(1 + at)^2 2\sum_{n=1}^{\infty} \frac{n(n+1)}{2}[at/1+at]^{n-1} - \sum_{n=1}^{\infty} n[at/1+at]^{n-1}$$

NOTES

$$= 1/(1 + at)^2 \left[2[1 - (at/1+at)]^{-3} - [1 - (at/1+at)]^{-2} \right]$$

$$\left(\text{since } (1- x)^{-3} = \sum_{n=1}^{\infty} \frac{n(n+1)}{2} x^{n-1}\right)$$

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

$$E(X^2(t)) = 1 + 2at$$

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

$$= 1 + 2at - 1 = 2at$$

since E(X(t)) and V(X(t)) are functions of t, {X(t)} is not stationary.

Example 2: Examine whether the poisson process {X(t)}, given by the probability law $P\{X(t) = r\} = \frac{e^{-\lambda t} (\lambda t)^r}{r!}$, r = 0, 1, …. is covariance stationary.

Solution:

The probability distribution of X(t) is a poisson distribution with parameter λt.
Therefore E{X(t)} = λt ≠ a constant.
Therefore the poisson process is not covariance stationary.

Example 3 : If {X(t)} ia wide sense stationary process with autocorrelation

$R(\tau) = Ae^{-\alpha|\tau|}$, determine the second-order moment of random variable X(8) – X(5).

Solution:

Second-order moment of random variable X(8) – X(5) is given by

$E[\{X(8) - X(5)\}^2] = E\{X^2(8)\} + E\{X^2(5)\} - 2E\{X(8)X(5)\}$.

Given $R(\tau) = Ae^{-\alpha|\tau|}$

i.e., $R(t_1, t_2) = Ae^{-\alpha|t_1 - t_2|}$

Therefore $E(X^2(t)) = R(t, t) = A$

Therefore $E\{X^2(8)\} = E\{X^2(5)\} = A$

Also $E\{X(8)\, X(5)\} = R(8, 5) = Ae^{-3\alpha}$

Therefore $E[\{X(8) - X(5)\}^2] = A + A - 2\, Ae^{-3\alpha}$

$E[\{X(8) - X(5)\}^2] = 2A(1 - e^{-3\alpha})$

NOTES

Example 4: If $X(t) = Y\cos\omega t + Z\sin\omega t$, where Y and Z are two independent normal random variables with mean zero and same SDs and ω is a constant, prove that $\{X(t)\}$ is a SSS process of order 2.

Solution:

Given $E(X) = E(Y) = 0$ and

$\sigma_Y = \sigma_Z = \sigma$ (say)

Therefore $Var(Y) = Var(Z) = \sigma^2$

Since X(t) is a linear combination of Y and Z, that are independent, X(t) follows a normal distribution with

$E[X(t)] = \cos\omega t E(Y) + \sin\omega t\, E(Z) = 0$

$Var[X(t)] = \cos^2\omega t \times E(Y^2) + \sin^2\omega t \times E(Z^2)$

Since $X(t_1)$ and $X(t_2)$ are each $N(0, \sigma)$, $X(t_1)$ and $X(t_2)$ are jointly normal with the joint pdf given by $f(x_1, x_2, t_1, t_2) =$

$$\frac{1}{2\pi\sigma^2 v(1-r^2)} \quad \exp[-(x_1^2 - 2rx_1x_2 + x_2^2)/2(1-r^2)\sigma^2],\ -\infty < x_1, x_2 < \infty$$

where r = correlation coefficient between $X(t_1)$ and $X(t_2)$

$$= \frac{1}{\sigma^2} E[X(t_1) \cdot X(t_2)]$$

$$= \frac{1}{\sigma} E(Y\cos\omega t_1 + Z\sin\omega t_1)(Y\cos\omega t_2 + Z\sin\omega t_2)$$

$$= \frac{1}{\sigma^2} [E(Y^2)\cos\omega t_1 \cos\omega t_2 + E(Z^2)\sin\omega t_1 \sin\omega t_2)$$

$(E(YZ) = E(Y)E(Z) = 0)$

$$= \frac{1}{\sigma^2} [\sigma^2\cos\omega t_1 \cos\omega t_2 + \sigma^2\sin\omega t_1 \sin\omega t_2)$$

$$= \cos\omega(t_1 - t_2)$$

In a similar manner the joint pdf of $X(t_1 + h)$ and $X(t_2 + h)$ can be expressed with

$$R = \cos\omega((t_1 + h) - (t_2 + h))$$

$$= \cos\omega(t_1 - t_2)$$

Since the joint pdf of $\{X(t_1), X(t_2)\}$ and $\{X(t_1 + h), X(t_2 + h)\}$ are the same, $\{X(t)\}$ is a SSS of order 2.

NOTES

Example 5: Show that the process

$X(t) = A\cos\lambda t + B\sin\lambda t$ (where A & B are random variables) is wide- sense stationary, if

i)$E(A) = E(B) = 0$

ii) $E(A^2) = E(B^2)$ and

iii) $E(AB) = 0$.

Solution:

$E(X(t)) = \cos\lambda t \times E(A) + \sin\lambda t \times E(B)$ (1)

If $\{X(t)\}$is to be a WSS process, $E(X(t))$ must be a constant.(i.e independent of t).

In (1) if E(A) and E(B) are any constants other than zero, $E\{X(t)\}$ will be a function of t.

Therefore $E(A) = E(B) = 0$

$R(t_1, t_2) = E\{X(t_1) \times X(t_2)\}$

$= E\{(A\cos\lambda t_1 + B\sin\lambda t_1)(A\cos\lambda t_2 + B\sin\lambda t_2)\}$

$= E(A^2)\cos\lambda t_1 \cos\lambda t_2 + B^2 \sin\lambda t_1 \sin\lambda t_2 + E(AB)\sin\lambda(t_1 + t_2)$ (2)

If $\{X(t)\}$ is to be a WSS process, $R(t_1, t_2)$ must be a function of t_1-t_2.

Therefore in (2) $E(AB) = 0$ and $E(A^2) = E(B^2) = u$

Then $R(t_1, t_2) = u\cos\lambda(t_1 - t_2)$.

Example 6: Consider a random process $X(t) = A\cos(\omega_0 t + \varphi)$Where A and φ are independent random variables and φ is uniformly distributed in the interval $-\pi$ to π. Find the First and second moment of the process.

Solution:

As the random variables A and φ are independent

$f_{A\varphi}(a,) = f_A(a).\, f_\varphi(\varphi)$

φ is uniformly distributed in the interval $-\pi$ to π

$f_\varphi(\varphi) = 1/2\pi, \; -\pi < \varphi < \pi$

$f_{A\varphi}(a,) = f_A(a)\,(1/2\pi)$

NOTES

First moment

$$E(X(t)) = \int_{-\infty}^{\infty}\int_{-\pi}^{\pi} A\cos(\omega_0 t+\varphi)\, f_A(a)\,(1/2\pi)\,da\,d\varphi$$

$$= \int_{-\infty}^{\infty} A\, f_A(a).da\,(1/2\pi)\int_{-\pi}^{\pi}\cos(\omega_0 t+\varphi)d\varphi$$

$$= \int_{-\infty}^{\infty} A\, f_A(a).da\,(1/2\pi)\,.\,0 = 0 \qquad \left[\int_{-\pi}^{\pi}\cos(\omega_0 t+\varphi)d\varphi = 0\right]$$

First moment = 0

Second moment is

$$E(X^2(t)) = \int_{-\infty}^{\infty}\int_{-\pi}^{\pi} A^2\cos^2(\omega_0 t+\varphi)\, f_A(a)\,(1/2\pi)\,da\,d\varphi$$

$$= \int_{-\infty}^{\infty} A^2\, f_A(a).da\,(1/2\pi)\int_{-\pi}^{\pi}\cos^2(\omega_0 t+\varphi)d\varphi$$

The first integral is the mean square value of the random variable A.

$$\text{So, } E(X^2(t)) = \overline{A^2}\,(1/2\pi)\int_{-\pi}^{\pi}\frac{(1+\cos 2(\omega_0 t+\varphi))}{2}d\varphi \qquad \left[\cos^2\theta = \frac{1+\cos 2\theta}{2}\right]$$

$$= \overline{A^2}\,(1/2\pi)\,½\,.\,2\pi$$

$$= ½\,\overline{A^2}$$

Example 7: For a random process X(t) = Y sin ωt, Y is a uniform random variable in the interval -1 to 1. Check whether the process is wide sense stationary or not..

Solution:

To prove wide- sense stationary we have to prove the its mean is a constant and auto correlation depends only on the time difference.

As Y is uniformly distributed in the interval -1 to 1

$f_Y(y) = ½ , -1 < y < 1$

Mean of the process is

$$E\{X(t_1)\} = \int_{-1}^{1} y \sin\omega t_1\, f_Y(y)dy$$

NOTES

$$= \int_{-1}^{1} y \sin \omega t_1 \, \tfrac{1}{2} \, dy$$

$$= \sin \omega t_1 \, \tfrac{1}{2} \int_{-1}^{1} y \, dy$$

$$= 0$$

The autocorrelation of the process is

$$E\{X(t_1) \times X(t_2)\} = E(Y \sin \omega t_1 \, Y \sin \omega t_2)$$

$$= \sin \omega t_1 \sin \omega t_2 \, E(Y^2)$$

$$= \sin \omega t_1 \sin \omega t_2 \int_{-1}^{1} y^2 f_Y(y) dy$$

$$= \sin \omega t_1 \sin \omega t_2 \int_{-1}^{1} y^2 (1/2) dy$$

$$= \sin \omega t_1 \sin \omega t_2 (1/3)$$

$$= (1/3) \frac{\cos(t_1 - t_2) - \cos(t_1 + t_2)}{2}$$

Though the first moment namely mean is constant, the autocorrelation function is not a function of time difference of the two random variables alone. So the process is not wide sense stationary.

Try yourself !

1. For the sine wave process $X(t) = Y \cos\omega_0 t$, $-8 < t < 8$, ω_0 = constant, the amplitude Y is a random variable with uniform distribution in the interval 0 to 1. Check whether the process is stationary or not.

2. Show that the random process $X(t) = A \cos(\omega_0 t + \varphi)$ is wide sense stationary if A and ω_0 are constants and φ is uniformly distributed random variable in $(0, 2\pi)$

3. Given a random variable with characteristic function $\varphi(\omega)$ and a random process $X(t) = \cos(\lambda t + Y)$. Show that $\{X(t)\}$ is stationary in the wide sense $\varphi(1) = 0$ and $\varphi(2) = 0$.

4.6 MARKOV PROCESS AND MARKOV CHAIN

A random process or stochastic process X(t) is said to be a Markov process if given the value of X(t), the value X(v) for $v > t$ does not depend on the values of X(u) for $u < t$. In other words, the future behavior of the process depends only on the present value and not on the past values.

NOTES

4.6.1 Markov Process

A random process X(t) is said to be Markovian if

$P[X(t_{n+1}) \leq x_{n+1}/X(t_n) = x_n, X(t_{n-1}) = x_{n-1}, ..X(t_0) = x_0] = P[X(t_{n+1}) \leq x_{n+1} / X(t_n) = x_n]$

Where $t_0 \leq t_1 \leq t_2 \leq \ldots\ldots \leq t_n \leq t_{n+1}$.

$X_0, X_1, X_2, \ldots X_n, X_{n+1}$ are called the states of the process.

If the random process at time at time t_n is in the state X_n, the future state of the random process X_{n+1} at time t_{n+1} depends only on the present state X_n and not on the past states $X_{n-1}, X_{n-2}, \ldots X_0$.

Examples of Markov process.

1.The probability of raining today depends on previous weather conditions existed for the last two days and not on past weather conditions.

2.A first order linear difference equation is Markovian.

4.6.1.1 Classification of Markov process.

A Markov process can be classified into four types based on the nature of the values taken by t and $\{X_i\}$.

i) A continuous random process satisfying Markov property is called a continuous parameter Markov process as t and $\{X_i\}$ are both continuous.

ii) A continuous random sequence satisfying Markov property is called a discrete parameter Markov process as the parameter t is discrete but $\{X_i\}$ is continuous.

iii) A discrete random process satisfying Markov property is called a continuous parameter Markov chain as t is continuous and $\{X_i\}$is discrete.

iv) A discrete random sequence satisfying Markov property is called a continuous parameter Markov chain as t is discrete and $\{X_i\}$ is also discrete.

4.6.2 Markov Chain

If for all n, $P[X_n = a_n / X_{n-1} = a_{n-1}, X_{n-2} = a_{n-2}, \ldots X_0 = a_0] = P[X_n = a_n / X_{n-1} = a_{n-1}]$, then the process $\{X_n\}$, n= 0,1,2, … is called Markov chain.

$(a_1, a_2, ..a_n, \ldots)$ are called the states of the Markov chain.

The joint probability of Markov chain is

$P[X_n = a_n, X_{n-1} = a_{n-1}, X_{n-2} = a_{n-2}, \ldots X_0 = a_0] = P[X_n = a_n / X_{n-1} = a_{n-1}]$

$P[X_n = a_n / X_{n-1} = a_{n-1}] \ldots\ldots P[X_1 = a_1 / X_0 = a_0] P[X_0 = a_0]$

NOTES

$$= P[X_0 = a_0] \prod_{m=1}^{n} P[X_m = a_m / X_{m-1} = a_{m-1}]$$

The conditional probability $P[X_n = a_j / X_{n-1} = a_i]$ is called the **one-step transition probability** from state a_i to state a_j at the nth step and is denoted by $p_{ij}(n-1, n)$.

If the one-step transition probability does not depend on the step

i.e., $p_{ij}(n-1, n) = p_{ij}(m-1, m)$ the Markov chain is called a **Homogeneous Markov chain** or the chain is said to have stationary transition probabilities.

When the Markov chain is homogeneous, the one-step transition probability is denoted p^{ij}. The matrix $P = \{p_{ij}\}$ is called (one-step) **transition probability matrix (t p m).**
Note: The tpm of a Markov chain is a stochastic matrix, since $p_{ij} \geq 0$ and $\sum_j p_{ij} = 1$

for all i, i.e., the sum of all the elements of any row of the tpm is 1. This is obvious because the transition from state a_i to any state, including a_i is a certain event.

The conditional probability that the process is in state a_j at step n given that it was in a state a_i at step 0, i.e, $P[X_n = a_j / X_0 = a_i]$ is called the **n-step transition probability** and denoted by $p_{ij}^{(n)}$.

$$p_{ij}^{(1)} = p_{ij}$$

Let us consider an example in which we explain how the tpm is formed fro a Markov chain. Assume that a man is at an integral point of the x-axis between the origin and the point x = 3. He takes a unit step either to the right with probability 0.7 or to the left with probability 0.3, unless he is at the origin when he takes a step to the right to each x = 1 or he is at the point x = 3, when he takes a step to the left to reach x = 2. The chain is called 'Random walk with reflecting barriers'

The tpm is given below

States of X_{n-1} \ States of X_n	0	1	2	3
0	0	1	0	0
1	0.3	0	0.7	0
2	0	0	0	0.7
3	0	0	1	0

Note: p_{23} = the element in the 2nd row, 3rd row column of this tpm = 0.7. This means that, if the process is at state 2 at step (n -1), the probability that it moves to state 3 at step n = 0.7, where n is any positive integer.

NOTES

If the probability that the process is in state a_i is $p_i (i = 1,2,\ldots k)$ at any arbitrary step, then the row vector $p = (p_1, p_2, \ldots . p_k)$ is called the probability distribution of the process at that time. In particular $P^{(0)} = \{p_1^{(0)}, p_2^{(0)}, \ldots . . p_k^{(0)}\}$ is the initial probability distribution,

where $p_1^{(0)} + p_2^{(0)} + \ldots . . + p_k^{(0)} = 1$.

[Remark: The transition probability matrix together with the initial probability distribution completely specifies a Markov chain $\{X_n\}$. In the example given above, let us assume the initial probability distribution of the chain is

$P^{(0)} = (¼, ¼, ¼, ¼)$

i.e., $P\{X_0 = i\} = ¼, i = 0,1,2,3$

Then we have, for example given below

$P\{X_1 = 2 / X_0 = 1\} = 0.7$: $P\{X_2 = 1/X_1 = 2\} = 0.3$,

$P\{X_2 = 1, X_1 = 2 / X_0 = 1\} = P\{X_2 = 1 / X_1 = 2\} \times P\{X_1 = 2/ X_0 = 1\} = 0.3 \times 0.7 = 0.21$

$P\{X_2 = 1, X_1 = 2 , X_0 = 1\} = P\{ X_0 = 1\} \times P\{X_2 = 1, X_1 = 2/X_0 = 1\} = ¼ \times 0.21 = 0.0525$

$P\{X_3 = 3, X_2 = 1, X_1 = 2, X_0 = 1\} = P\{X_2 = 1, X_1 = 2, X_0 = 1\} \times P\{X_3 = 3 / X_2 = 1, X_1 = 2, X_0 = 1\} = 0.0525 \times 0 = 0$

4.5.2.1Chapman-Kolmogorov Theorem

If P is the tpm of a homogeneous Markov chain, then n-step tpm $P^{(n)}$ is equal to P^n.

i.e.,$[P_{ij}^{(n)}] = [P_{ij}]^n$.

If $p = \{p_i\}$ is the state probability distribution of the process at an arbitrary time, then that after one step is pP, where P is the tpm of the chain and that after n step is pP^n.

A stochastic matrix P is said to be a **regular** matrix, if all the entries of p^m (for some positive integer m) are positive. A homogeneous Markov chain is said to be regular if its tpm is regular.

If a homogeneous Markov chain is regular, then every sequence of state probability distributions approaches a unique fixed probability distribution called the **stationary distribution** or steady-state distribution of the Markov chain.

i.e., $\lim_{n \to \infty} [p^{(n)}] = \pi$, where the state probability distribution at step n,

$P^{(n)} = \{p_1^{(n)}, p_2^{(n)}, \ldots . . p_k^{(n)}\}$ and the stationary distribution $\pi = (\pi_1, \pi_2, \ldots \pi_k)$ are row vectors.

NOTES

Note:

If P is the tpm of the regular chain, then $\pi P = \pi$ (π is a row vector).

4.6.2.2 Classification of states of a Markov chain:

1.State j is said to be **accessible** from state i if $P_{ij}^{(n)} > 0$ for some n 24 0.

2.Two states i and j which are accessible to each other are said to **communicate**

i) state i communicates with itself for all i 24 0.

ii) If state i communicates with state j and state j communicates with state k, then state i communicates with state k.

3.Two states that communicate are in the same **class**. Two classes of state are either identical or disjoint.

If $P_{ij}^{(n)} > 0$ for some n and for all i and j, then every state can be reached from every other state. When this condition is satisfied, the Markov chain is said to be **irreducible**.

The tpm of an irreducible chain is an irreducible matrix. Otherwise the chain is said to be non-irreducible or irreducible.

A state is said to be an **absorbing state** if no other state is accessible from it; that is, for an absorbing state i, $p_{ii} = 1$.

State i of a Markov chain is called a return state, if $p_{ii}^{(n)} > 0$ for some $n > 1$.

The period di of a return state i is defined as the greatest common divisor of all m such that

$p^{ii(m)} > 0$, i.e.,$d_i = GCD\{m: p_{ii}^{(m)} > 0\}$. State i said to be **periodic** with period d_i

if $d_i > 1$ and **aperiodic** if $d_i = 1$.

State i is aperiodic if p_{ii} 4 4 0. The probability that the chain returns to state i, having started from state i, for the first time at the nth step(or after n transitions) is denoted by

$f_{ii}^{(n)}$ and called the first return time probability or the recurrence time probability.

$\{n, f_{ii}^{(n)}\}$, n = 1,2,3, …. is the distribution of recurrence times of the state i.

If $F_{ii} = \sum_{n=1}^{\infty} f_{ii}^{(n)} = 1$, the return to state i is certain.

$\mu_{ii} = \sum_{n=1}^{\infty} n f_{ii}^{(n)}$ is called the mean recurrence time of the state i.

A state i is said to be **persistent** or **recurrent** if the return to state i is certain, i.e., if $F_{ii} = 1$.

NOTES

The State i is said to be transient if the return to state i is uncertain, i.e., if $F_{ii} < 1$. The state i is said to be non-null persistent if its mean recurrence time μ_{ii} is finite and null persistent if $\mu_{ii} = \infty$.

A non-null persistent and aperiodic state is called **ergodic.**

Theorem: If a Markov chain is irreducible, all its states are of the same type. They are all transient, all null persistent or all non-null persistent. All its states are either aperiodic or periodic with same period.

Theorem: If a Markov chain is finite irreducible, all its states are non-null persistent

Example 1: A raining process is considered as a two state Markov chain. If it rains, it is considered to be in state 0 and if does not rain, the chain is in state 1. The transition probability of Markov chain is defined as

$$P = \begin{pmatrix} 0.6 & 0.4 \\ 0.2 & 0.8 \end{pmatrix}$$

Find the probability that it will rain for three days from today assuming that it is raining today. Find also the unconditional probability that it will rain after three days. Assume the initial probabilities of state 0 and1 as 0.4 and 0.6 respectively.

Solution:

The one-step transition probability matrix is given as

$$P = \begin{pmatrix} 0.6 & 0.4 \\ 0.2 & 0.8 \end{pmatrix}$$

$$P^2 = \begin{pmatrix} 0.44 & 0.56 \\ 0.28 & 0.72 \end{pmatrix}$$

$$P^3 = \begin{pmatrix} 0.376 & 0.624 \\ 0.312 & 0.688 \end{pmatrix}$$

Therefore the probability that it will rain on third day given that it will rain today is 0.376

The unconditional probability that it will rain after three days is

$P(X_3 = 0) = 0.4P_{00}^{3} + 0.6P_{10}^{3} = (0.4)(0.376) + (0.6)(0.312) = 0.3376.$

Example 2: A person owning a scooter has the option to switch over to scooter, bike or a car next time with the probability of (0.3, 0.5, 0.2). If the transition probability matrix is

$$\begin{pmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.5 & 0.3 \\ 0.25 & 0.25 & 0.5 \end{pmatrix}$$

What are the probabilities related to his fourth purchase?

NOTES

Solution:

$$P = \begin{pmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.5 & 0.3 \\ 0.25 & 0.25 & 0.5 \end{pmatrix}$$

$$P^3 = \begin{pmatrix} 0.277 & 0.351 & 0.372 \\ 0.269 & 0.359 & 0.372 \\ 0.275 & 0.345 & 0.380 \end{pmatrix}$$

Probabilities of his fourth purchase = $(0.3, 0.5, 0.2)P^3 = (0.2726, 0.3538, 0.3736)$

Example 3: The transition probability matrix of a Markov chain $\{X_n\}$, n= 1,2,…having 3 states 1,2 and 3 is

$$P = \begin{pmatrix} 0.1 & 0.5 & 0.4 \\ 0.6 & 0.2 & 0.2 \\ 0.3 & 0.4 & 0.3 \end{pmatrix}$$

and the initial distribution $p(0) = (0.7, 0.2, 0.1)$

Find i)$P\{X_2 =3\}$ and ii) $P\{X_3 = 2, X_2 = 3, X_1 = 3, X_0 = 2\}$.

Solution:

$$P = \begin{pmatrix} 0.1 & 0.5 & 0.4 \\ 0.6 & 0.2 & 0.2 \\ 0.3 & 0.4 & 0.3 \end{pmatrix}$$

$$P^2 = \begin{pmatrix} 0.43 & 0.31 & 0.26 \\ 0.24 & 0.42 & 0.34 \\ 0.36 & 0.35 & 0.29 \end{pmatrix}$$

i) $P\{X_2 = 3\} = \sum_{i=1}^{3} P\{X_2 = 3/X_0 = i\} \times P\{X_0 = i\}$

$= P\{X_2 = 3/X_0 = 1\} \times P\{X_0 = 1\} + P\{X_2 = 3/X_0 = 2\} \times P\{X_0 = 2\} +$

$P\{X_2 = 3/X_0 = 3\} \times P\{X_0 = 3\}$

$= p_{13}^{(2)} \times 0.7 + p_{23}^{(2)} \times 0.2 + p_{33}^{(2)} \times 0.1$

$= 0.26 \times 0.7 + 0.34 \times 0.2 + 0.29 \times 0.1$

$= 0.279$

ii) $P\{X_3 = 2, X_2 = 3, X_1 = 3, X_0 = 2\}$

$= P\{X_3 = 2/ X_2 = 3, X_1 = 3, X_0 = 2\} \times P\{ X_2 = 3, X_1 = 3, X_0 = 2\}$.(by Markov property)

NOTES

$$= P\{X_3 = 2/ X_2 = 3\} \times P\{ X_2 = 3/ X_1 = 3, X_0 = 2\} \times P\{X_1 = 3, X_0 = 2\}$$

$$= P\{X_3 = 2/ X_2 = 3\} \times P\{ X_2 = 3/ X_1 = 3\} \times P\{X_1 = 3/X_0 = 2\} \times P\{X_0 = 2\}$$

$$= p_{32} \times p_{33} \times p_{23} \times 0.2 = 0.4 \times 0.3 \times 0.2 \times 0.2 = 0.0048.$$

Example 3: Using limiting behaviour of Homogeneous chain, find the steady state probabilities of the chain given by the transition matrix

$$\text{Probability } P = \begin{pmatrix} 0.1 & 0.6 & 0.3 \\ 0.5 & 0.1 & 0.4 \\ 0.1 & 0.2 & 0.7 \end{pmatrix}$$

Solution:

We know that, if P is the tpm of the regular chain, then $\pi P = \pi$ and $\sum \pi = 1$

$$[\pi_1 \ \pi_2 \ \pi_3] = [\pi_1 \ \pi_2 \ \pi_3] \begin{pmatrix} 0.1 & 0.6 & 0.3 \\ 0.5 & 0.1 & 0.4 \\ 0.1 & 0.2 & 0.7 \end{pmatrix}$$

$$\pi_1 = 0.1\,\pi_1 + 0.5\,\pi_2 + 0.1\,\pi_3$$

$$\pi_2 = 0.6\,\pi_1 + 0.1\,\pi_2 + 0.2\,\pi_3$$

$$\pi_3 = 0.3\,\pi_3 + 0.4\,\pi_2 + 0.7\,\pi_3$$

$$\pi_1 + \pi_2 + \pi_3 = 1$$

Solving these equations

$\pi_1 = 0.2021$

$\pi_2 = 0.2553$

$\pi_3 = 0.5426$

The steady state probability is (0.2021, 0.2553, 0.5426)

Example 4: Find the nature of the states of the Markov chain with tpm

$$P = \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{pmatrix} 0 & 1 & 0 \\ ½ & 0 & ½ \\ 0 & 1 & 0 \end{pmatrix} \end{array}$$

Solution:

$$\text{Given } P = \begin{pmatrix} 0 & 1 & 0 \\ ½ & 0 & ½ \\ 0 & 1 & 0 \end{pmatrix}$$

NOTES

$$P^2 = \begin{pmatrix} ½ & 0 & ½ \\ 0 & 1 & 0 \\ ½ & 0 & ½ \end{pmatrix}$$

$P = P^3 : P^4 = P^2$.

In general $P^{2n} = P^2$

We note that $p_{00}^{(2)} > 0$; $p_{01}^{(1)} > 0$; $p_{02}^{(2)} > 0$

$p_{10}^{(1)} > 0$; $p_{11}^{(2)} > 0$; $p_{12}^{(1)} > 0$

$p_{20}^{(2)} > 0$; $p_{21}^{(1)} > 0$; $p_{22}^{(2)} > 0$

Therefore the Markov chain is irreducible.

Also $p_{ii}^{(2)} = p_{ii}^{(4)} = p_{ii}^{(6)} = \ldots\ldots = p_{ii}^{(2n)} = \ldots\ldots > 0$, for all i, all the states a re periodic and the period $= GCD\{2,4,6 \ldots .\} = 2$

That is the period is 2.

Therefore the chain is finite and irreducible, all its states are non-null persistent. All the states are not ergodic.

Example 5: A man tosses a fair coin until 3 heads occur in a row. Let X_n denotes the longest string of heads ending at the nth trial. Show that the process is Markovian. Find the transition matrix and classify the states.

Solution: The state space = {0, 1, 2, 3}, since the coin is tossed until 3 heads occur in a row.

The transition probability matrix is

$$P = \begin{array}{c} \\ X_{n-1} \\ \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{cccc} \multicolumn{4}{c}{X_n} \\ 0 & 1 & 2 & 3 \end{array} \\ \begin{pmatrix} ½ & ½ & 0 & 0 \\ ½ & 0 & ½ & 0 \\ ½ & 0 & ½ & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{array} \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \end{array}$$

Solution:

Given

$$P = \begin{pmatrix} ½ & ½ & 0 & 0 \\ ½ & 0 & ½ & 0 \\ ½ & 0 & ½ & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

NOTES

$$P^2 = \begin{pmatrix} ½ & ¼ & ¼ & 0 \\ ½ & ¼ & ¼ & 0 \\ ½ & ¼ & ¼ & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

$$P^3 = \begin{pmatrix} ½ & ¼ & ¼ & 0 \\ ½ & ¼ & ¼ & 0 \\ ½ & ¼ & ¼ & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

It is clear that the chain is irreducible.

State 3 is absorbing.

$P_{ii}^{(2)} = p_{ii}^{(3)} = p_{ii}^{(4)} = \ldots\ldots > 0$ for all i

$d_i = GCD\{2,3,4, \ldots\ldots\} = 1$

Therefore all the states are aperiodic.

Example 6: Consider a Markov chain with state space {0, 1} and the tpm

$$P = \begin{pmatrix} 1 & 0 \\ ½ & ½ \end{pmatrix}$$

i) Draw a transition diagram.

ii) Show that state 0 is recurrent.

iii) Show that state 1 is transient.

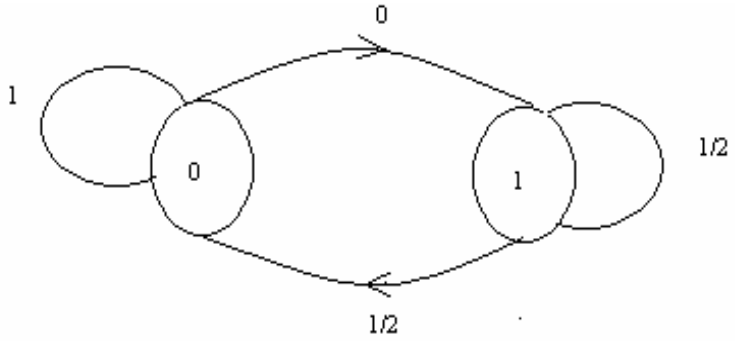
iv) Is the state 1 periodic? If so what is the priod?

v) Is the chain irreducible.

vi) Is the chain ergodic? Explain.

Solution:

i)



NOTES

Given

$$P = \begin{pmatrix} 1 & 0 \\ ½ & ½ \end{pmatrix}$$

$$P^2 = \begin{pmatrix} 1 & 0 \\ ¾ & ¼ \end{pmatrix}$$

$$P^3 = \begin{pmatrix} 1 & 0 \\ 7/8 & 1/80 \end{pmatrix}$$

$$P^4 = \begin{pmatrix} 1 & 0 \\ 15/16 & 1/16 \end{pmatrix}$$

ii) A state i is said to be recurrent if and only if starting from state i, the process eventually returns to state i with probability one.
Hence by the definition of recurrent, state 0 is recurrent.

i) A sate i is said to be transient if and only if there is a +ve probability that the process will not return to this state.

ii) We have

$f_{11}^{(1)} = ½ > 0$; $f_{11}^{(2)} = ¼ > 0$

GCD (1, 2, 3, ...) = 1

Hence the state 1 is periodic with period 1, i.e., aperiodic.

iii) The chain is not irreducible as $p_{01}^{(n)} = 0$, n = 1,2,3,. . . ..

iv) Since the chain is not irreducible, all the states are not non-null persistent. Hence the chain is not ergodic.

Example 7: A fair die is tossed repeatedly. If X_n denotes the maximum of the numbers occurring in the first n tosses, find the transition probability matrix P of the Markov chain $\{X_n\}$. Find also P^2 and $P(X_2 = 6)$

Solution:

The state space is given by {1, 2, 3, 4, 5, 6}.

Let X_n = maximum of the numbers obtained in the first n trials = 4 (say)

Then X_{n+1} = 4 if (n + 1)th trial is 1, 2, 3 or 4.
= 5 if (n + 1)th trial is 5.
= 6 if (n + 1)th trial is 6.

$P(X_{n+1} = 4 / X_n = 4) = 1/6 + 1/6 + 1/6 + 1/6 = 2/3$

$P(X_{n+1} = i / X_n = 4) = 1/6$, i = 5,6.

NOTES

Therefore the transition probability matrix is given by

$$P = \begin{pmatrix} 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 0 & 2/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 0 & 0 & 3/6 & 1/6 & 1/6 & 1/6 \\ 0 & 0 & 0 & 4/6 & 1/6 & 1/6 \\ 0 & 0 & 0 & 0 & 5/6 & 1/6 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

$$P^2 = \frac{1}{36} \begin{pmatrix} 1 & 3 & 5 & 7 & 9 & 11 \\ 0 & 4 & 5 & 7 & 9 & 11 \\ 0 & 0 & 9 & 7 & 9 & 11 \\ 0 & 0 & 0 & 16 & 9 & 11 \\ 0 & 0 & 0 & 0 & 25 & 11 \\ 0 & 0 & 0 & 0 & 0 & 36 \end{pmatrix}$$

Initial state probabilities are $P(X_i = 0) = P(0) = 1/6, i = 1, \ldots 6$

Now $P(X_2 = 6) = \sum_{i=1}^{6} P(X_2 = 6 / X_0 = i)\, P(X_0 = i)$

$$= 1/6 \sum_{i=1}^{6} p_{i6}^{\,2}$$

$$= 1/6 \times 1/36(11 + 11 + 11 + 11 + 11 + 36)$$

$$= 91/26$$

Example 8: There are two white marbles in urn A and 3 red marbles in urn B. At each step of he process, a marble is selected from each urn and the two marbles selected are interchanged. Let the state ai of the system be the number of red marbles in A after 3 steps? In the long run, what is the probability that there are 2 red marbles in urn A?

NOTES

Solution:

State space of the chain $\{X_n\} = \{0, 1, 2\}$, since the number of balls in the urn A is always 2.

Let the tpm of the chain $\{X_n\}$ be

$$P = \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{pmatrix} p_{00} & p_{01} & p_{02} \\ p_{10} & p_{11} & p_{12} \\ p_{20} & p_{21} & p_{22} \end{pmatrix} \end{array}$$

$p_{00} = 0$ (since the state cannot remain at 0 after interchange of marbles)

$p_{02} = p_{20} = 0$ (Since the number of red marbles in urn cannot increase or decrease by 2 in one interchange)

To start with, A contains 0 red marble. After an interchange, A will contain1 red marble (and 1 white marble) certainly.

Therefore $p_{01} = 1$.

Let $X_n = 1$, i.e., A contains 1 red marble (and 1 white marble) and B contains 1 white and 2 red marbles.

Then $X_{n+1} = 0$, if A contains 0 red marble (and 2 white marbles) and B contains 3 red marbles. i.e., if 1 red marble is chosen from A and 1 white marble is chosen from B and interchanged.

Therefore $P[X_{n+1} = 0 / X_n = 1] = p_{10} = ½ \times 1/3 = 1/6$

$p_{12} = ½ \times 2/3 = 1/3$

we know that $p_{10} + p_{11} + p_{12} = 1$

Therefore $p_{11} = 1-(p_{10} + p_{12}) = ½$

$p_{21} = 2/3$

$p_{22} = 1-(p_{20} + p_{21}) = 1/3$

Therefore the tpm is

$$P = \begin{pmatrix} 0 & 1 & 0 \\ 1/6 & ½ & 1/3 \\ 0 & 2/3 & 1/3 \end{pmatrix}$$

Now p(0) = (1, 0, 0) as there is no red marble in A in the beginning

$p^{(1)} = p^{(0)}P = (0, 1, 0)$

NOTES

$p^{(2)} = p^{(1)}P = (1/6, ½, 1/3)$

$p^{(3)} = p^{(2)}P = (1/12, 23/26, 5/18)$

Therefore P[there are 2 red marbles in A after 3 steps] = $P[X_3 = 2] = 5/18$.

Let the stationary probability distribution of the chain be $\pi = (\pi_0, \pi_1, \pi_2)$

We know that, if P is the tpm of the regular chain, then $\pi P = \pi$ and ? $\pi = 1$

$$(\pi_0, \pi_1, \pi_2) \begin{pmatrix} 0 & 1 & 0 \\ 1/6 & ½ & 1/3 \\ 0 & 2/3 & 1/3 \end{pmatrix} = (\pi_0, \pi_1, \pi_2)$$

$1/6\ \pi_1 = \pi_0$
$\pi_0 + ½\ \pi_1 + 2/3\ \pi_2 = \pi_1$
$1/3\ \pi_1 + 1/3\ \pi_2 = \pi_2$
$\pi_0 + \pi_1 + \pi_2 = 1$

Solving these equations we get

$\pi_0 = 1/10,\ \pi_1 = 6/10,\ \pi_2 = 3/10$

P[there are 2 red marbles in A in the long run] = 0.3

Example 9: Three boys A,B and C are throwing a ball to each other. A always throw the ball to B and B always throws the ball to C, but C is just as likely to throw the ball to B as to A. Show that the process is Markovian. Find the transition matrix and classify the states.

Solution: The transition probability matrix of the process $\{X_n\}$ is given below

$$P = \begin{array}{c} \\ A \\ B \\ C \end{array} \begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ ½ & ½ & 0 \end{pmatrix} \end{array} \quad \begin{array}{l} X_n \\ X_{n-1} \end{array}$$

States X_n depend only on states of Xn-1, but not on earlier states.

Therefore $\{X_n\}$ is markovian chain.

NOTES

$$P^2 = \begin{pmatrix} 0 & 0 & 1 \\ ½ & ½ & 0 \\ 0 & ½ & ½ \end{pmatrix}$$

$$P^3 = \begin{pmatrix} ½ & ½ & 0 \\ 0 & ½ & ½ \\ ¼ & ¼ & ½ \end{pmatrix}$$

$p_{11}^{(3)} > 0$, $p_{12}^{(1)} > 0$, $p_{13}^{(2)} > 0$, $p_{21}^{(2)} > 0$, $p_{22}^{(2)} > 0$, $p_{23}^{(1)} > 0$, $p_{31}^{(1)} > 0$, $p_{32}^{(1)} > 0$, $p_{33}^{(2)} > 0$.
Therefore the chain is irreducible.

$$P^4 = \begin{pmatrix} 0 & ½ & ½ \\ ¼ & ¼ & ½ \\ ¼ & ½ & ¼ \end{pmatrix}$$

$$P^5 = \begin{pmatrix} ¼ & ¼ & ½ \\ ¼ & ½ & ¼ \\ 1/8 & 3/8 & ½ \end{pmatrix}$$

$$P^6 = \begin{pmatrix} ¼ & ½ & ¼ \\ ¼ & 3/8 & ½ \\ 1/8 & 3/8 & 3/8 \end{pmatrix}$$

and so on.

Now $p_{ii}^{(2)} > 0$, $p_{ii}^{(3)} > 0$, $p_{ii}^{(5)} > 0$, $p_{ii}^{(6)} > 0$ etc for i = 2,3, and GCD(2, 3, 5, 6, ...) = 1
Therefore the states 2 and 3 are periodic. i.e., the states A and B are periodic with period 1 i.e., aperiodic.

Now $p_{11}^{(3)} > 0$, $p_{11}^{(5)} > 0$, $p_{11}^{(6)} > 0$ etc and GCD(3,5,6, ...) = 1
Therefore the state 1 is periodic with period 1 i.e., aperiodic.

Since the chain is finite and irreducible, all its states and non-null persistent.
Moreover all the states are ergodic.

Try yourself !

1) The tpm of a Markov chain {Xn}, n = 1,2,3, . . . with 3 states 0, 1, & 2 is

$$P = \begin{pmatrix} ¾ & ¼ & 0 \\ ¼ & ½ & ¼ \\ 0 & ¾ & ¼ \end{pmatrix}$$

NOTES

with initial distributions p(0) = (1/3, 1/3, 1/3).
Find i)$P(X_2 = 2)$ ii)$P(X_3 = 1, X_2 = 2, X_1 = 2, X_0 = 2)$
(**Solution:** i) 1/6 ii) 3/64)

2) The one-step transition probability matrix of a Markovian chain with sate {0, 1} is given as

$$P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

a)draw a transition diagram. b) Is it irreducible Markov chain?

3) Suppose that the probability of a dry day (state 0) following a rainy day (state 1) is 1/3 and that the probability of a rainy day following a dry day is ½ . Given that May 1 is a dry day, find the probability that i) May 3 is also a dry day ii) May 5 is also a dry day.
(**Solution:** i)5/12 ii) 173/432)

4) A housewife buys 3 kinds of cereals, A, B and C. She never buys the same cereal in successive weeks. If she buys cereal A, the next wee she buys cereal B. However if she buys B or C the next week she is 3 times as likely as to buy A as the other cereal. How often she buys each of the three cereals.
(**Solution :** (-3/7, 16/35 , 4/35))

4.7 POISSON PROCESS

In this section we consider an important example of a discrete random process kown as the Poisson process.

4.7.1 Definition: Let X(t) represents the number of occurrences of a certain event in (0, t) then the discrete random process {X(t): t = 0} is a Poisson process provided the following postulates are satisfied

i) $P(1 \text{ occurrence in } (t, t + \Delta t)) = \lambda \Delta t + 0(\Delta t)$
ii) $P(0 \text{ occurrence in } (t, t + \Delta t)) = 1 - \lambda \Delta t + 0(\Delta t)$
iii) $P(2 \text{ or more in } (t, t + \Delta t)) = 0(\Delta t)$
iv) X(t) is independent of the number of occurrence of the event in any interval prior or after (0, t).
v) The probability that the event occurs a specified number of times in $(t_0, t_0 + t)$ depends only on t but not on t_0.

Let the number of events in the interval [0, t] is denoted by X(t). Then the stochastic process {X(t): t = 0} is a Poisson process, with mean λ. Note that the number of events in the interval [0, t] is a Poisson distribution with parameter λt.

NOTES

The probability distribution of X(t) is given by

$$P(X(t) = n) = \frac{e^{-\lambda t}(\lambda t)^n}{n!}, \quad n = 0, 1, 2, \ldots .$$

4.7.2 Second-order probability function of a homogeneous Poisson process.

$P[X(t_1) = n_1, X(t_2) = n_2] = P[X(t_1) = n_1]P[X(t_2) = n_2 / X(t_1) = n_1]$ $t_2 > t_1$

$= P[X(t_1) = n_1]$ P[the event occurs $(n_2 - n_1)$ times in the interval length $(t_2 - t_1)$]

$$= \frac{e^{-\lambda t_1}(\lambda t_1)^{n_1}}{n_1!} \quad \frac{e^{-\lambda(t_2 - t_1)}[\lambda(t_2 - t_1)]^{(n_2 - n_1)}}{(n_2 - n_1)!}, \quad \text{if } n_2 \overset{24}{12} n_1$$

$$= \begin{cases} \dfrac{e^{-\lambda t_2}\lambda^{n_2} t_1^{n_1} (t_2 - t_1)^{(n_2 - n_1)}}{n_1!\ (n_2 - n_1)!} & \text{if } n_2 \overset{24}{12} n_1 \\ 0 \text{ otherwise} \end{cases}$$

Proceeding similarly, we can get the third-order probability function as

$P[X(t_1) = n_1, X(t_2) = n_2, , X(t_3) = n_3]$

$$= \begin{cases} \dfrac{e^{-\lambda t_3}\lambda^{n_3} t_1^{n_1} (t_2 - t_1)^{(n_2 - n_1)} (t_3 - t_2)^{(n_3 - n_2)}}{n_1!\ (n_2 - n_1)!(n_3 - n_2)!} & \text{if } n_3 \overset{24}{12} n_2 \overset{24}{12} n_1 \\ 0 \text{ otherwise} \end{cases}$$

4.7.3 Mean and Variance of the Poisson process.

The probability law of the Poisson process {X(t)} is the same as that of a Poisson distribution with parameter λt.

$E\{X(t)\} = Var\{X(t)\} = \lambda t$.

4.7.4 Autocorrelation of the Poisson processs

$E(X^2(t)) = \lambda t + \lambda^2 t^2$.

$R_{xx}(t_1, t_2) = E[X(t_1) X(t_2)]$

$= E[X(t_1) \{X(t_2) - X(t_1) + X(t_1)\}]$

$= E[X(t_1) \{X(t_2) - X(t_1)\}] + E\{X^2(t_1)\}$

$= E[X(t_1)]E[X(t_2) - X(t_1)] + E\{X^2(t_1)\}$

since {X(t)} is a process of independent increments.

$= \lambda t_1[\lambda(t_2 - t_1)] + \lambda t_1 + \lambda^2 t_1^2$, if $t_2 \overset{24}{12} t_1$

$= \lambda^2 t_1 t_2 + \lambda t_1$, if $t_2 \overset{24}{12} t_1$

$R_{xx}(t_1, t_2) = \lambda^2 t_1 t_2 + \lambda t_1$, if $t_2 \overset{24}{12} t_1$

NOTES

4.7.5 Auto covariance of the Poissson process

$$C_{xx}(t_1, t_2) = R_{xx}(t_1, t_2) - E[X(t_1)]\, E[X(t_2)]$$
$$= \lambda^2 t_1 t_2 + \lambda t_1 - \lambda^2 t_1 t_2$$
$$= \lambda t_1, \text{ if } t_2 \geq t_1$$

4.7.6 Correlation coefficient

$$r_{xx}(t_1, t_2) = \frac{C_{xx}(t_1, t_2)}{\sqrt{\text{Var}[X(t_1)]\text{Var}[X(t_2)]}} = \frac{\lambda t_1}{\sqrt{(\lambda t_1\, \lambda t_2)}} = \sqrt{(t_1/t_2)}, \text{ if } t_2 \geq t_1$$

4.7.7 Properties of Poisson process

1. *The Poisson process is not a stationary process.*

The probability distribution of X(t) is a Poisson distribution with parameter λt.
Therefore $E\{X(t)\} = \lambda t \geq 4$ a constant.
Therefore the Poisson process is not covariance stationary.

2. *The Poisson process is a Markov process.*

Proof: Consider $P[X(t_3) = n_3 / X(t_2) = n_2, X(t_1) = n_1]$

$$= \frac{P[X(t_1) = n_1, X(t_2) = n_2, X(t_3) = n_3]}{P[X(t_1) = n_1, X(t_2) = n_2]}$$

$$= \frac{e^{-\lambda(t_3 - t_2)} \lambda^{(n_3 - n_2)} (t_3 - t_2)^{(n_3 - n_2)}}{(n_3 - n_2)!}$$

(Refer section 4.4.2)

$$= P[X(t_3) = n_3 / X(t_2) = n_2]$$

This means that the conditional probability distribution of distribution of $X(t_3)$ given all the past values $X(t_1) = n_1, X(t_2) = n_2$ depends only on the most recent value $X(t_2) = n_2$. That is, the Poisson process possesses the Markov property. Hence the result.

3. *The sum of two Poisson process is a Poisson process.*

Let $\{X_1(t): t \geq 0\}$, $\{X_2(t): t \geq 0\}$ be 2 Poisson process and let $X(t) = X_1(t) + X_2(t)$.

$$P(X(t) = n) = \sum_{r=0}^{n} P(X_1(t) = r)\, P(X_2(t) = n - r)$$

$$= \sum_{r=0}^{n} \frac{e^{-\lambda_1 t}(\lambda_1 t)^r}{r!} \cdot \frac{e^{-\lambda_2 t}[\lambda(t_2)]^{(n-r)}}{(n - r)!}, \text{ if } n_2 \geq n_1$$

NOTES

$$= \frac{e^{-(\lambda_1 + \lambda_2)t} \; nCr(\lambda_1 t)^r(\lambda_2 t)^{n-r}}{n!}$$

$$= e^{-(\lambda_1 + \lambda_2)t} \frac{[(\lambda_1 + \lambda_2)t]^n}{n!}$$

which is a Poisson process with parameter $(\lambda_1 + \lambda_2)t$.

4. *The difference of two Poisson process is not a Poisson process.*

Proof: Let $X(t) = X_1(t) - X_2(t)$.

$E(X(t)) = E[X_1(t) - X_2(t)]$

$= E[X_1(t)] - E[X_2(t)]$

$E(X^2(t)) = E[X_1(t) - X_2(t)]^2$

$= E[X_1^2(t)] - E[X_2^2(t)] - 2E[X_1(t)X_2(t)]$

$= (\lambda_1 t + \lambda_1^2 t^2) + (\lambda_2 t + \lambda_2^2 t^2) - 2(\lambda_1 t)(\lambda_2 t)$

$= t^2(\lambda_1^2 + \lambda_2^2 - 2\lambda_1\lambda_2) + (\lambda_1 + \lambda_2)t$

$= t^2(\lambda_1 - \lambda_2)^2 + (\lambda_1 + \lambda_2)t$

? $t^2(\lambda_1 - \lambda_2)^2 + (\lambda_1 - \lambda_2)t$

where $E(X_2(t))$ for a Poisson process $X(t)$ with parameters λ is given by $\lambda t + \lambda^2 t^2$. Therefore $X(t)$ is a Poisson process.

5. *The interarrival time of a Poisson process, i.e., the interval between two successive occurrences of a Poisson process with parameter λ has an exponential distribution with mean $1/\lambda$.*

Proof:

Let E_i and E_{i+1} two consecutive events.

Let E_i take place at the time instant t_i and T be the interval between the occurrences of E_i and E_{i+1}.T is a continuous random variable.

$P(T > t)$ = P{ E_{i+1} did not occur in (ti, ti + t)}

= P{No event occurs in an interval of length t}

= P[X(t) = 0]

= $e^{-\lambda t}$

Therefore the cdf of T is given by

$F(T) = P\{T \overset{c}{\lambda} t\} = 1 - e^{-\lambda t}$

Therefore the pdf of T is given by

$f(t) = \lambda e^{-\lambda t}$ (t $\overset{24}{12}$ 0)

which is an exponential distribution with mean $1/\lambda$.

NOTES

6. *If the number of occurrences of an event E is an interval of length t is a Poisson process {X(t)} with parameter λ and each occurrence of E has a constant probability p of being recorded and the recordings are independent of each other, then the number N(t) of the recorded occurrences in t is also a Poisson process with parameter λp.*

Proof:

$$P[N(t) = n] = \sum_{r=0}^{\infty} P\{E \text{ occurs } (n + r) \text{ times in t and n of them are recorded}\}$$

$$= \sum_{r=0}^{\infty} \frac{e^{-\lambda t}(\lambda t)^{n+r}}{(n + r)!} \quad (n + r)C_n p^n q^r, \quad q = 1 - p$$

$$= \sum_{r=0}^{\infty} \frac{e^{-\lambda t}(\lambda t)^{n+r}}{(n + r)!} \quad \frac{(n + r)!\, p^n q^r}{n!\ r!}$$

$$= \frac{e^{-\lambda t}(\lambda pt)^n}{n!} \sum_{r=0}^{\infty} \frac{(\lambda qt)^r}{r!}$$

$$= \frac{e^{-\lambda t}(\lambda pt)^n\ e^{\lambda qt}}{n!}$$

$$= \frac{e^{-\lambda pt}(\lambda pt)^n}{n!}$$

Example 1: Derive the probability law for the Poisson process.

Solution:

Let λ be the number of occurrences of the event in unit time.

Let $P_n(t) = P\{X(t) = n\}$

Therefore $P_n(t + \Delta t) = P\{X(t + \Delta t) = n\}$

$= P\{(n - 1)$ calls in $(0, t)$ and 1 call in $(t, t + \Delta t)\} + P\{n$ calls in $(0, t)$ and no calls in $(t, t + \Delta t)\}$

$= P_{n-1}(t)\, \lambda \Delta t + P_n(t)\, (1 - \lambda \Delta t)$

$$\frac{P_n(t + \Delta t) - P_n(t)}{\Delta t} = \lambda\{P_{n-1}(t) - P_n(t)\}$$

Taking the limits as $\Delta t \to 0$

$$\frac{d}{dt} P_n(t) = \lambda\{P_{n-1}(t) - P_n(t)\} \qquad (1)$$

Let the solution of the equation (1) be

$$P_n(t) = \frac{(\lambda t)^n f(t)}{n!} \qquad (2)$$

Differentiating (2) with respect to t,

$$P'_n(t) = \lambda^n \{nt^{n-1} f(t) + t^n f'(t)\} \qquad (3)$$

NOTES

Using (2) and (3) in (1)

$$\frac{\lambda^n t^n f^{|}(t)}{n!} = \frac{-\lambda (\lambda t)^n f(t)}{n!}$$

i.e., $f^{|}(t) = -\lambda f(t)$

$$f(t) = ke^{-\lambda t} \qquad (4)$$

From (2), $f(0) = P_0(0) = P[X(0) = 0]$
$= P\{\text{no event occurs in } (0, 0)\}$
$= 1.$ (5)

Using (5) in (4), we get k = 1 and hence

$$f(t) = e^{-\lambda t} \qquad (6)$$

Using (6) in (2),

$$Pn(t) = P(X(t) = n) = \frac{e^{-\lambda t}(\lambda t)^n}{n!}, \; n = 0, 1, 2, \ldots .$$

Thus probability distribution of X(t) is given by

$$P(X(t) = n) = \frac{e^{-\lambda t}(\lambda t)^n}{n!}, \; n = 0, 1, 2, \ldots .$$

Example 2: If $\{X_1(t)\}$ and $\{X_2(t)\}$ are two independent Poisson processes with parameters λ_1 and λ_2 respectively, show that

$$P[X_1(t) = k / \{X_1(t) + X_2(t) = n\}] = {}^nC_k \, p^k q^{n-k}$$

Where $p = \frac{\lambda_1}{\lambda_1 + \lambda_2}$ and $q = \frac{\lambda_2}{\lambda_1 + \lambda_2}$.

Solution:

Required conditional probability

$$= \frac{P[\{X_1(t) = k\} \cap \{X_1(t) + X_2(t) = n\}]}{P\{X_1(t) + X_2(t) = n\}}$$

$$= \frac{P[\{X_1(t) = k\} \cap \{X_2(t) = n - k\}]}{P\{X_1(t) + X_2(t) = n\}}$$

$$= \frac{e^{-\lambda_1 t}(\lambda_1 t)^k}{k!} \times \frac{e^{-\lambda_2 t}(\lambda_2 t)^{n-k}}{(n-k)!} \Big/ \frac{e^{-(\lambda_1 + \lambda_2)t}\{(\lambda_1 + \lambda_2)t\}^n}{n!}$$

(by independence and additive property)

$$= \frac{n!}{k!\,(n-k)!} \frac{(\lambda_1 t)^k (\lambda_2 t)^{n-k}}{\{(\lambda_1 + \lambda_2)t\}^n}$$

NOTES

$$= nC_k \left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^k \left(\frac{\lambda_2}{\lambda_1+\lambda_2}\right)^{n-k}$$

$$= nC_k\, p^k q^{n-k}$$

Where $p = \frac{\lambda_1}{\lambda_1+\lambda_2}$ and $q = \frac{\lambda_2}{\lambda_1+\lambda_2}$.

Example 3: If **{X(**t): t ε T**}** is a Poisson process then the auto correlation coefficient between X(t) and X(t + s) is $\left(\frac{t}{(t+s)}\right)^{½}$.

Solution:

If **{X(**t): t ε T**}** be a Poisson process with parameter λ. Then

$E(X(t)) = \lambda t$, $Var(X(t)) = \lambda t$

$E(X^2(t)) = Var(X(t)) + [E(X(t)]^2$

$= \lambda t + (\lambda t)^2$

By definition auto correlation is

$$\rho(t, t+s) = \frac{C(t, t+s)}{[Var(X(t))\, Var(X(t+s))]^{½}}$$

$C(t, t+s) = E[X(t)\, X(t+s)] - E(X(t))E(X(t+s))$

Consider

$E(X(t))E(X(t+s)) = E[(X(t))(X(t+s)) - X(t) + X(t)]$

$= E(X(t)X(t)) + E(X(t)\, X(t+s) - X(t))$

$= E(X(t)^2) + E(X(t))E\,(X(t+s) - X(t))$

$= E(X^2(t)) + E(X(t))E\,(X(t+s) - E(X(t)^2)$

$= \lambda t + (\lambda t)^2 + \lambda t.\lambda s$

Substituting this in C(t, t + s) we have

$C(t, t+s) = \lambda t + (\lambda t)^2 + \lambda t.\lambda s - \lambda t.\lambda(t+s)$

$= \lambda t$

$$\rho(t, t+s) = \frac{\lambda t}{[\lambda t.\lambda(t+s)]^{½}}$$

$$= \frac{t}{(t^2+ts)^{½}}$$

$$= \left(\frac{t}{(t+s)}\right)^{½}$$

Example 3: Suppose the customers are arriving at a ticket counter according to a Poisson process with a mean rate of 2 per minute. Then in an arrival of 5 minutes find the probability that the number of customers arriving is i) exactly 3 ii) greater than 3 iii) less than 3.

NOTES

Solution:

Given $\lambda = 2$

Therefore the number of customers X(t) arriving in an interval of duration t minutes follows a Poisson distribution with mean 2.

i)The probability the number of customers arriving is exactly 3 is

$$P[X(5) = 3] = \frac{e^{-10}(10)^3}{3!}$$

ii)The probability the number of customers arriving is greater than 3 is

$$P[X(5) > 3] = 1 - P[X(5) \leq 3]$$

$$= 1 - \sum_{k=0}^{3} \frac{e^{-10}(10)^k}{k!}$$

iii)The probability the number of customers arriving is less than 3 is

$$P[X(5) < 3] = \sum_{k=0}^{2} \frac{e^{-10}(10)^k}{k!}$$

Example 4: A machine goes out of order, whenever a component fails. The failure of this part follows a Poisson process with a mean rate of 1 per week. Find the probability that 2 weeks have elapsed since last failure. If there are 5 spare parts of this component in an inventory and that the ext supply is not due in 10 weeks, find the probability that the machine will not be out of order in the next 10 weeks.

Solution:

Here the unit time is 1 week.

Mean failure rate = mean number of failures in a week = $\lambda = 1$

P(no failures in he2 weeks since last failure) = P[X(2) = 0]

$$= \frac{e^{-\lambda t}(\lambda t)^n}{n!} = \frac{e^{-2}(2)^0}{0!} = e^{-2} = 0.135$$

There are only 5 spare parts and the machine should not go out of order in the next weeks

$$P[X(10) \leq 5] = \sum_{n=0}^{5} \frac{e^{-10}\,10^n}{n!} = 0.068s$$

Example 5: If customers arrive at a customer in accordance with a mean rate of 2 per minute, find the probability that the arrivals is i) more than 1 minute ii) between 1 minute and 2 minute and iii) 4 minute or less.

NOTES

Solution:
The interval T between 2 consecutive arrivals follows an exponential distribution with parameter $\lambda = 2$.

$$P(T > 1) = \int_1^{\infty} 2 e^{-t} dt = e^{-2} = 0.135$$

i) $$P(1 < T < 2) = \int_1^2 2 e^{-2t} dt = e^{-2} - e^{-4} = 0.177$$

iii) $$P(T \overset{c}{\lambda} 4) = \int_0^4 2 e^{-2t} dt = 1 - e^{-8} = 0.999$$

Example 6: A radioactive source emits a particle at a rate of 5 per minute in accordance with Poisson process. Each particle emitted has a probability 0.6 of being recorded. Find the probability that 10 particles are recorded in 4-min period.

Solution: We know that
If the number of occurrences of an event E is an interval of length t is a Poisson process {X(t)} with parameter λ and each occurrence of E has a constant probability p of being recorded and the recordings are independent of each other, then the number N(t) of the recorded occurrences in t is also a Poisson process with parameter λp.

Here $\lambda = 5$ and $p = 0.6$

$$P(N(t) = k) = \frac{e^{-3t} (3t)^k}{k!}$$

$$P(N(4) = 10) = \frac{e^{-12} (12)^{10}}{10!} = 0.014$$

Example 7: The number of accidents in a city follows a Poisson process with a mean of 2 per day and the number X_i of people involved in the i th accident has the distribution (independent)
$P\{X_i = k\} = 1/2^k$ (k = 1). Find the mean and variance of the number of people involved in accidents per week.

Solution:
The mean and variance of the distribution $P\{X_i = k\} = 1/2^k$, $k = 1, 2, 3, \ldots . \infty$ can be obtained as 2 and 2.

Let the number of accidents on any day be assumed as n.
The numbers of people involved in these accidents be $X_1, X_2, X_3, \ldots . X_n$

NOTES

$X_1, X_2, X_3, \ldots . X_n$ are independent and identically distributed random variables with mean 2 and variance 2.

Therefore, by central limit theorem $(X_1 + X_2 + X_3 + \ldots + X_n)$ follows a normal distribution with mean 2n and variance 2n, i.e,. the total number of people involved in all the accidents on a day with n accidents = 2n.

If N denotes the number of people involved in accidents on any day, then
P(N = 2n) = P[X(t)) = n] where X(t) is the number of accidents.

$$= \frac{e^{-2t}(2t)^n}{2!}$$

$$E(N) = \sum_{n=0}^{\infty} \frac{2n\, e^{-2t}\,(2t)^n}{n!}$$

$$= 2E\{X(t)\} = 4t$$

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

$$= \sum_{n=0}^{\infty} \frac{4n^2 e^{-2t}\,(2t)^n}{n!} - 16t^2$$

$$= 4E\{X^2(t)\} - 16t^2$$

$$= 4\{Var(X(t)\} + [E(X(t))]^2 - 16t^2$$

$$= 4[2t + 4t^2] - 16t^2 = 8t$$

Therefore, mean and variance of the number of people involved in accidents per week are 28 and 56 respectively.

Example 8: If {X(t)} is a Poisson process, prove that

$P\{X(s) = r/\ X(t) = n\} = nCr\ (s/t)^r (1 - s/t)^{n-r}$ where $s < t$

Solution:

$$P\{X(s) = r/\ X(t) = n\} = \frac{P\{X(s) = r\}1\{X(t) = n\}}{P\{X(t) = n\}}$$

$$= \frac{P\{X(s) = r\}1\{X(t - s) = n - r\}}{P\{X(t) = n\}}$$

$$= \frac{P\{X(s) = r\}P\{X(t - s) = n - r\}}{P\{X(t) = n\}} \text{ (by independence)}$$

$$= \frac{e^{-\lambda s}(\lambda s)^r}{r!} \times \frac{e^{-\lambda(t-s)}(\lambda(t - s))^{n-r}}{(n - r)!}$$

$$\frac{e^{-\lambda t}\ (\lambda t)^n}{n!}$$

NOTES

$$= \frac{n!}{r!(n-r)!} \frac{s^r (t-s)^{n-r}}{t^n}$$

$$= nCr\ (s/t)^r (1 - s/t)^{n-r}$$

Example 9: If the particles are emitted from a radio active source at the rate of 20 per hour. Find the probability that exactly 5 particles are emitted during 15 minute period.

Solution:
Let X(t) denote the number of particles emitted during t minutes. By Poisson process

$P(X(t) = n) = \frac{e^{-\lambda t}(\lambda t)^n}{n!}$, n = 0, 1, 2, ...

here $\lambda = 20/60 = 1/3$

$P(X(t) = n) = \frac{e^{-t/3}(t/3)^n}{n!}$, n = 0, 1, 2, ...

P[exactly 5 particles are emitted during a 15 minute period]

$= P[X(15) = 5]$

$= \frac{e^{-5}(5)^5}{5!}$

$= 0.1755$

Example 10: The probability that a person is suffering from cancer is 0.001. Find the probability that out of 4000 persons a) exactly 4 suffer because of cancer, b) more than 3 persons will suffer from the disease.

Solution:
The probability that a person is suffering from cancer = p = 0.001
No. of persons = 4000
Mean = $\lambda = np = 4000 \times 0.001 = 4$

By the definition of Poisson distribution

$P[X = x] = \frac{e^{-\lambda}\lambda^x}{x!}$, x = 0, 1, 2, ...

a) $P[X = 4] = \frac{e^{-4}4^4}{4!} = 0.19532$

b) $P[X > 3] = 1 - P[X = 3] = 1 - \sum_{x=0}^{3} \frac{e^{-4}4^x}{x!}$

$= 1 - 0.4333 = 0.5666$

Try yourself !

NOTES

1. A radio active source emits particles at a rate 6 per minute in accordance with Poisson process.
   Each particle emitted has a probability 1/3 of being recorded. Find the probability that atleast 5 particles are recorded in a 5 minute period.
   (**Solution:** 0.9707)

2. If patients arrive at a clinic according to Poisson process with mean rate of 2 per minute. Find the probability that during a 1-minute interval, no patient arrives.
   (**Solution:** 0.135)

3. Suppose that customers arrive at a bank according to a Poisson process with a mean rate of 3 per minute; find the probability that during a time interval of 2 minutes i)exactly 4 customers arrive and ii)more than 4 customers arrive.
   (**Solution**: 0.133, 0.715)

4. Assume that the number of messages input to a communication channel in an interval of duration t seconds is a Poisson process with mean rate $\lambda = 0.3$. Compute i)the probability that exactly three messages will arrive during a ten-second interval. ii) The probability that the number of message arrivals ina interval of duration five seconds is between three and seven.
   (**Solution:** 0.224, 0.191)

How you understood ?

Say True

1. For a first-order stationary process, the density function of different random variables are different.
2. A wide-sense stationary process is a second-ordinary process.
3. A transition probability matrix of a finite sate Markov chain is a square matrix.
4. A stationary transition matrix of a finite state Markov chain always implies a stationary random sequence.
5. The transition probability matrix of a finite state Markov chain takes only non-negative values.
6. Poisson process is a discrete random process.
7. Poisson process is stationary
8. The interarrival time of a Poisson is also Poisson..
9. Poisson process is Markovian.
10. The number of rows in a transition probability matrix of a finite state Markov chain is equal to the number of states of the system.

(Answers: 1.false, 2.false, 3.true, 4.false, 5.true, 6.true, 7.false, 8.false, 9.true, 10.true)

NOTES

Short answer questions

1. What is a random process?
2. Classify Random process. Give example for each.
3. What is a first ordinary stationary process?
4. What is a wide-sense stationary process? Give an example.
5. Define a strict sense stationary process.
6. Classify Markov process.
7. Define Markov chain.
8. Define absorbing state, recurrent state and transient state of a Markov chain.
9. Give the properties of Poisson process.
10. Why a Poisson process is non-stationary ?

REFERENCES:

1. T.Veerarajan, “Probability, statistics and Random Process”, Tata McGraw Hill, 2002.
2. P.Kandasamy, K. Thilagavathi and K. Gunavathi, “Probability, Random Variables and Random processors”, S. Chand, 2003.

NOTES

UNIT 5

QUEUEING THEORY

- **Introduction**
- **Single and multi-server Markovian Queues**
- **Little's Formula**
- **Average measures**

5.1 INTRODUCTION

Queuing or waiting lines arises in many situations of our daily life. For example passengers waiting for ticket booking ii) Machines wait for repair iii) patients waiting for treatment etc are different forms of queue being formed. If the queue is very long and customers in that have no patience to wait for service, they will seek other outlet. In such situations the servicing center may incur in loss. Therefore to solve the above problem the study of queuing theory is very important. Queuing theory analysis involves the study of systems behavior overtime.

5.2 LEARNING OBJECTIVES

The students

- will be exposed to basic characteristic features of a queuing system and acquire skills in analyzing queuing models.
- Will have a well founded knowledge of different queuing models which can describe real life phenomena.

5.3 BASIC CHARACTERISTIC OF QUEUEING PHENOMENA

1) Input or arrival (inter-arrival distribution)
2) Output or departure (service) distribution
3) Service channels
4) Service discipline
5) Maximum number of customers allowed in the system
6) Calling service or population.

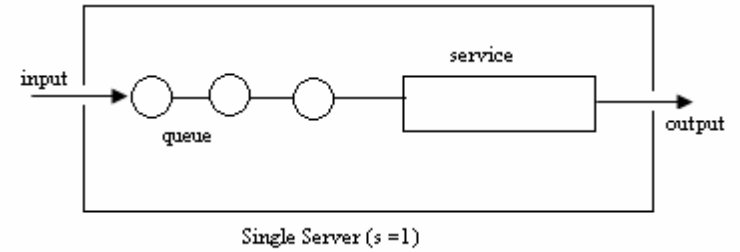
NOTES

- *Customer*: The units which arrive at a service centre are called customers.
- *Queue* (waiting line): number of customers to be serviced
- *Queuing system:* Number of customers in waiting line and also customers being serviced.
- *Service channel*: the system that renders service to customers. The service channel may be single (one unit) or multi channel (more than one).

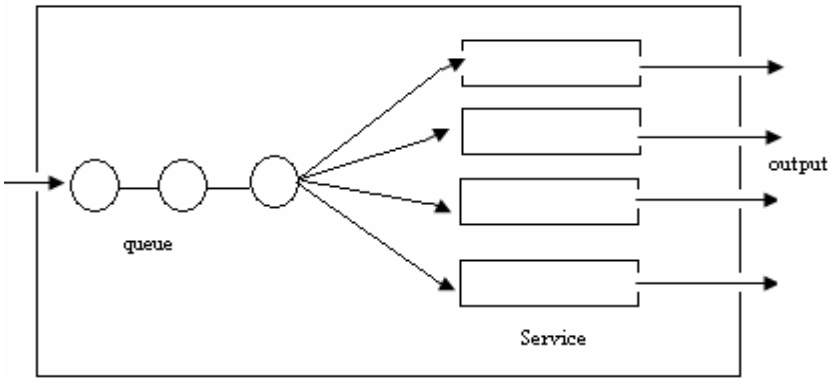
There may be one or more service stations or facilities at the service centre. If the service station is free, the customer who arrives there will be served immediately. If not the customer has to wait in line for his turn.

Types of queuing models.

There are several types of queuing models. Some of them are
1.Single queue – single service point.
2.Multiple queues – multiple service points
3.Simple queue – multiple service points
4.Multiple queues – single service point.
The most common case of queuing models is the single channel waiting line.

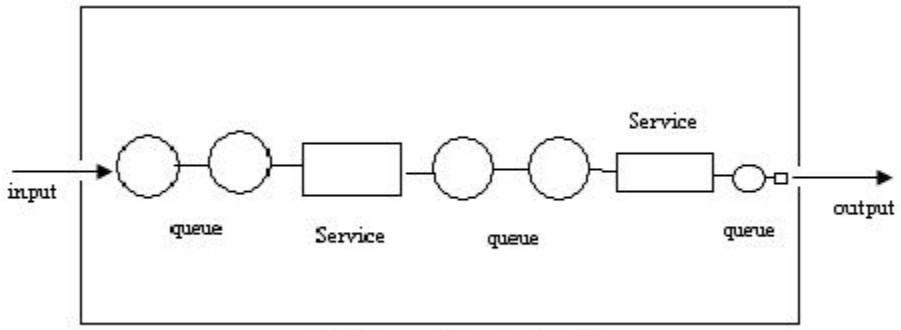


Single Server (s =1)



Multi Servers (in parallel) (s =4)

NOTES



Multi Servers (in series) (s = 3)

- *Arrival rate* (λ) is the number of customers arriving per unit of time.
- *Service rate* (μ) is the number of customers served per unit of time.
- *Service discipline:* is the order of service rule based on which customers in queue are serviced.

For example :

i) First come first served (FCFS) or First in first out (FIFO) as per which the customers are served in the strict order of their arrival.
ii) If the last arrival in the system is served first, we have Last come first served (LCFS)
iii) If the service is given in random order, Service in random order (SIRO) discipline.
iv) General service discipline (GD)

5.4 OPERATING CHARACTERISTICS OF QUEUEING SYSTEM

The operating characteristic of a queueing system refer to the numerical values of the probability distributions of various decision variables like arrival rate, number of facilities, service time, line length, priority system etc., some common characteristics area given below:

❖ Queue length:

Probability distribution of queue length can be obtained with the help of the given probability distribution of the arrival and service process. A large queue indicates poor service facility or a need for more space. On the other hand small queue indicates excess of service facilities.

❖ Waiting time in the queue:

It refers to the time spent by the customers in the queue before the commencement of his service. Long waiting time may increase the customer's dissatisfaction and potential loss of future revenues.

NOTES

❖ Waiting time in system

This is the total time spent by a customer in the queue plus service time. Long waiting time may indicate need for a change in the priority rules.

❖ State of the system

A basic concept in the analysis of a queueing theory is that of a state of the system. It involves study of a system's behavior overtime. It is classified as follows:

✓ Transient state

A queueing system is said to be in transient state when its operating characteristics are dependent on time. A queueing system is in transient system when the probability distributions of arrivals, waiting time and service time of the customers are dependent on time. This state occurs at the beginning of the operation of the system.

✓ Steady state

If the operating characteristics become independent of time, the queueing system is said to be in a steady state. Thus a queueing system acquires steady state when the probability distribution of arrivals, waiting time and servicing time of the customers are independent of time. This state occurs in the long run of the system.

5.5 KENDALL'S NOTATION FOR REPRESENTING QUEUEING MODELS

A general queueing system is denoted by **(a/b/c) : (d/e)** where

a = probability distribution of the inter-arrival time
b = probability distribution of the service time
c = number of servers in the system
d = maximum number of customers allowed in the system
e = queue discipline

Certain descriptive notations used for the arrivals and service time distributions i.e., to replace notation (a and b) are as follows.

M = inter-arrival times or service times having exponential distributions, where the letter M stands for '**Markovian**' property of the exponential.

D = inter-arrival times or service times that are constant or deterministic.

G = Service time distributions of general type i.e., no assumption is made about the form of distribution.

Notations and symbols:

The following symbols and notations will be used in connection with the queueing systems:

n = total number of customers in the system, both waiting and in service.

NOTES

λ = Average number of customers arriving per unit of time

μ = Average number of customers being served per unit of time

s = number of parallel service channels (servers)

$\mathbf{L_s}$ or $E(N_s) = E(n)$ = Expected or average number of customers in the system, both waiting and in service.

$\mathbf{L_q} = E(N_q) = E(m)$ = average or expected number of customers waiting in the queue (excluding those who are receiving the service)

$\mathbf{W_s} = E(W_s)$ = average or expected waiting time of a customer in the system both waiting and in service (including the service time).

$\mathbf{W_q} = E(W_q)$ = average or expected waiting time of a customer in the queue (excluding service time)

$\mathbf{P_n(t)}$ = probability that there are n customers in the system at any time t (both waiting and in service) assuming that the system has started its operation at time zero.

$\boldsymbol{\rho}$ = Traffic intensity or utilization factor which represents the proportion of time the servers are busy = $\boldsymbol{\lambda/\mu}$

$\mathbf{P_n}$ = steady state probability of having n customers in the system.

5.6 DIFFERENCE EQUATION RELATED TO POISSON QUEUE SYSTEM

Let $P_n(t)$ be the probability that there are n customers in the system at time t ($n > 0$). Let us first derive the differential equation satisfied by $P_n(t)$ and deduce the difference equation satisfied by P_n (probability of n customers at any time) in the steady-state.

Let λ_n be the average arrival rate when there are n customers in the system (both waiting in the queue and being served) and let μ_n be the average service rate when there are n customers in the system.

The system being in steady state does not mean that the arrival rate and service rate are independent of the number of customers in the system.

Now $P_n(t + \Delta t)$ is the probability of the n customers at time $t + \Delta t$.

The presence of n customers in the system at time $t + \Delta t$ can happen in any one of the following four mutually exclusive ways:

NOTES

i) Presence on n customers at t and no arrival or departure during Δt time.

ii) Presence of (n -1) customers at t and one arrival and no departure during Δt time.

iii) Presence of (n + 1) customers at t and no arrival and one departure during Δt time.

iv) Presence of n customers at t and one arrival and one departure during Δt time (since more than one arrival/departure during Δt time is ruled out)

Therefore

$$P_n(t+\Delta t) = P_n(t)(1-\lambda_n \Delta t)(1-\mu_n \Delta t) + P_{n-1}(t)\lambda_{n-1}\Delta t(1-\mu_n \Delta t) + P_{n+1}(t)(1-\lambda_{n+1}\Delta t)\mu_{n+1}\Delta t + P_n(t).\lambda_n \Delta t.\mu_n \Delta t$$

[since P(arrival occurs during $\Delta t = \lambda \Delta t$ etc)]

i.e., $P_n(t+\Delta t) = P_n(t) - (\lambda_n + \mu_n) P_n(t)\Delta t + \lambda_{n-1} P_{n-1}(t)\Delta t + \mu_{n+1} P_{n+1}(t)$, on omitting terms containing $(\Delta t)^2$ which is negligibly small.

$$\frac{P_n(t+\Delta t) - P_n(t)}{\Delta t} = \lambda_{n-1} P_{n-1}(t) - (\lambda_n + \mu_n) P_n(t)\Delta t + \mu_{n+1} P_{n+1}(t) \quad (1)$$

Taking the limits on both sides of (1) as $\Delta t \to 0$, we have

$$P_n'(t) = \lambda_{n-1} P_{n-1}(t) - (\lambda_n + \mu_n) P_n(t)\Delta t + \mu_{n+1} P_{n+1}(t) \quad (2)$$

Equation (2) does not hold good for n = 0, as $P_{n-1}(t)$ does not exist. Hence we derive the differential equation satisfied by $P_0(t)$ independently. Proceeding as before,

$P_0(t+\Delta t) = P_0(t))(1-\lambda_0 \Delta t) + P_1(t)(1-\lambda_1 \Delta t)\mu_1 \Delta t$, [by the possibilities (i) and (iii) given above and as no departure is possible when n = 0]

$$\frac{P_0(t+\Delta t) - P_0(t)}{\Delta t} = -\lambda_0 P_0(t) + \mu_1 P_1(t) \quad (3)$$

Taking limits on both sides of (3) as $\Delta t \to 0$, we have

$$P_0'(t) = -\lambda_0 P_0(t) + \mu_1 P_1(t) \quad (4)$$

Now in the steady state, $P_n(t)$and $P_0(t)$ are independent of time and hence $P_n'(t)$ and $P_0'(t)$ become zero. Hence the differential equations (2) and (4) reduce to the difference equations

$$\lambda_{n-1} P_{n-1} - (\lambda_n + \mu_n) P_n + \mu_{n+1} P_{n+1} = 0 \quad (5)$$

and $$\lambda_0 P_0 + \mu_1 P_1 = 0 \quad (6)$$

NOTES

Value of P_0 and P_n for Poisson Queue systems

From equation (6) derived above, we have

$$P_1 = \frac{\lambda_0}{\mu_1} P_0 \qquad (7)$$

Putting n =1 in (5) and using (7), we have

$$\mu_2 P_2 = (\lambda_1 + \mu_1) P_1 - \lambda_0 P_0$$
$$= (\lambda_1 + \mu_1) \frac{\lambda_0}{\mu_1} P_0 - \lambda_0 P_0 = \frac{\lambda_0 \lambda_1}{\mu_1} P_0$$

Therefore $P_2 = \frac{\lambda_0 \lambda_1}{\mu_1 \mu_2} P_0 \qquad (8)$

Successively putting n = 2,3,. . . in (5) and proceeding similarly, we can get

$$P_2 = \frac{\lambda_0 \lambda_1 \lambda_2}{\mu_1 \mu_2 \mu_3} P_0 \text{ etc.}$$

Finally $\mathbf{P_n = \frac{\lambda_0 \lambda_1 \lambda_2 \ldots \lambda_{n-1}}{\mu_1 \mu_2 \mu_3 \ldots \mu_n} P_0}$ **etc. for n = 1,2, ...** (9)

since the number of customers in the system can be 0 or 1 or 2 or 3 etc, which events are mutually exclusive and exhaustive, we have

$$\sum_{n=1}^{\infty} P_n$$

$$P_0 + \sum_{n=1}^{\infty} \left(\frac{\lambda_0 \lambda_1 \lambda_2 \ldots \lambda_{n-1}}{\mu_1 \mu_2 \mu_3 \ldots \mu_n} \right) P_0 = 1$$

Therefore $\mathbf{P_0 = \frac{1}{1 + \sum_{n=1}^{\infty} \left(\frac{\lambda_0 \lambda_1 \lambda_2 \ldots \lambda_{n-1}}{\mu_1 \mu_2 \mu_3 \ldots \mu_n} \right)}}$ (10)

Equations (9) and (10) will be used to derive the important characteristics of the four queueing models.

NOTES

5.7 CHARACTERISTICS OF INFINITE CAPACITY, SINGLE SERVER POISSON QUEUE MODEL I (M/M/1) : (∞/FIFO), when $\lambda_n = \lambda$ and $\mu_n = \mu$ ($\lambda < \mu$)

1.Average number L_s of customers in the system: Let N denote the number of customers in the queueing system (i.e., those in the queue and the one who is being served).

N is a discrete random variable, which can take the value 0,1,2, . . . ∞
Such that $P(N = n) = P_n = \left(\frac{\lambda}{\mu}\right)^n P_0$, from equation (9)

From equation (10), we have

$$\mathbf{P_0} = \frac{1}{1 + \sum_{n=1}^{\infty} \left(\frac{\lambda}{\mu}\right)^n} = \frac{1}{\sum_{n=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^n} = \mathbf{1 - \frac{\lambda}{\mu}}$$

$$\mathbf{P_n} = \left(\frac{\lambda}{\mu}\right)^n \left(\mathbf{1 - \frac{\lambda}{\mu}}\right)$$

Now $L_s = E(N) = \sum_{n=0}^{\infty} n \times P_n$

$$= \left(\frac{\lambda}{\mu}\right)\left(1 - \frac{\lambda}{\mu}\right) \sum_{n=1}^{\infty} n \left(\frac{\lambda}{\mu}\right)^{n-1}$$

$$= \frac{\lambda}{\mu}\left(1 - \frac{\lambda}{\mu}\right)\left(1 - \frac{\lambda}{\mu}\right)^{-2}, \text{ by binomial summation}$$

$$= \frac{\frac{\lambda}{\mu}}{1 - \left(\frac{\lambda}{\mu}\right)} = \frac{\lambda}{\mu - \lambda} \qquad (1)$$

2. Average number L_q of customers in the queue or average length of the queue:

If N is the number of customers in the system, then the number of customers in the queue is (N -1).

Therefore $L_q = E(N - 1) = \sum_{n=1}^{\infty} (n - 1) \times P_n$

$$= \left(1 - \frac{\lambda}{\mu}\right) \sum_{n=1}^{\infty} (n-1) \left(\frac{\lambda}{\mu}\right)^n$$

NOTES

$$= \left(\frac{\lambda}{\mu}\right)^2 1-\frac{\lambda}{\mu} \sum_{n=2}^{\infty} (n-1)\left(\frac{\lambda}{\mu}\right)^{n-2}$$

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

$$= \frac{\left(\frac{\lambda}{\mu}\right)^2}{1-\frac{\lambda}{\mu}} = \frac{\lambda^2}{\mu(\mu-\lambda)} \quad (2)$$

3. Average number L_w of customers in nonempty queues

$L_w = E\{(N-1)/(N-1) > 0\}$, since the queue is non-empty

$$= \frac{E(N-1)}{P(N-1>0)} = \frac{\lambda}{(\mu-\lambda)} \times \frac{1}{\sum_{n=2}^{\infty} P_n}$$

$$= \frac{\lambda^2}{\mu(\mu-\lambda)} \times \frac{1}{\sum_{n=2}^{\infty}\left(\frac{\lambda}{\mu}\right)^n\left(1-\frac{\lambda}{\mu}\right)}$$

$$= \frac{\lambda^2}{\mu(\mu-\lambda)} \times \frac{1}{\left(\frac{\lambda}{\mu}\right)^2\left(1-\frac{\lambda}{\mu}\right)\sum_{n=0}^{\infty}\left(\frac{\lambda}{\mu}\right)^n}$$

$$= \frac{\mu}{(\mu-\lambda)} \times \frac{1}{\left(1-\frac{\lambda}{\mu}\right)\left(1-\frac{\lambda}{\mu}\right)^{-1}}$$

$$= \frac{\mu}{(\mu-\lambda)} \quad (3)$$

4. Probability that the number of customers in the system exceeds k

$$P(N>k) = \sum_{n=k+1}^{\infty} P_n = \sum_{n=k+1}^{\infty} \frac{\lambda}{\mu}^n \; 1-\frac{\lambda}{\mu}$$

$$= \left(\frac{\lambda}{\mu}\right)^{k+1}\left(1-\frac{\lambda}{\mu}\right)\sum_{n=k+1}^{\infty}\left(\frac{\lambda}{\mu}\right)^{n-(k+1)}$$

NOTES

$$= \left(\frac{\lambda}{\mu}\right)^{k+1} \left(1 - \frac{\lambda}{\mu}\right) \sum_{n=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^{n}$$

$$= \left(\frac{\lambda}{\mu}\right)^{k+1} \left(1 - \frac{\lambda}{\mu}\right) \left(1 - \frac{\lambda}{\mu}\right)^{-1}$$

$$= \left(\frac{\lambda}{\mu}\right)^{k+1} \quad (4)$$

5.Probability density function of the waiting time in the system

Let W_s be the continuous random variable that represents the waiting time of a customer n the system, viz, the time between arrival and completion of service.

Let its pdf be f(*w*) and let f (*w*/n) be the density function of W_s subject to the condition that there are n customers in the queueing system when the customer arrives,

Then $f(w) = \sum_{n=0}^{\infty} f(w/n)\, P_n$ (5)

Now f (*w*/n) = pdf of sum of (n +1) service times (one part-service time of the customer

being served + n complete service times)

= pdf of sum of (n +1) independent random variables, each of which is exponentially distributed with parameter μ

$= \mu \frac{^{n+1}}{n!} e^{-\mu w} w^n$: $w > 0$ which is the pdf of Erlang distribution with parameter μ and n +1.

Therefore $f(w) = \sum_{n=0}^{\infty} \frac{\mu^{n+1}}{n!} e^{-\mu w} w^n \left(\frac{\lambda}{\mu}\right)^n \left(1 - \frac{\lambda}{\mu}\right)$ by (5)

$$= \mu e^{-\mu w} \left(1 - \frac{\lambda}{\mu}\right) \sum_{n=0}^{\infty} \frac{1}{n!} (\lambda w)^n$$

$$= \mu \left(1 - \frac{\lambda}{\mu}\right) e^{-\mu w} e^{\lambda w}, \text{ by exponential summation}$$

$$= (\mu - \lambda)\, e^{-\mu w} e^{-\lambda w}$$

$$= (\mu - \lambda)\, e^{-(\mu - \lambda) w} \quad (6)$$

which is the pdf of an exponential distribution with parameter (μ – ë).

NOTES

6.Average waiting time of a customer in the system.

W_s follows an exponential distribution with parameter $(\mu - \lambda)$.

$$E(W_s) = \frac{1}{\mu - \lambda} \qquad (7)$$

(since the mean of an exponential distribution is the reciprocal of its parameter)

7.Probability that the waiting time of a customer in the system exceeds t

$$P(W_s > t) = \int_t^\infty f(w)\, dw$$

$$= \int_t^\infty (\mu - \lambda)\, e^{-(\mu-\lambda)w}\, dw$$

$$= (\mu - \lambda)\, \frac{e^{-(\mu-\lambda)w}}{-(\mu-\lambda)} \Bigg|_t^\infty$$

$$= e^{-(\mu-\lambda)t} \qquad (8)$$

8. Probability density function of the waiting time in the queue

W_q represents the time between arrival and reach of service point.

Let pdf of W_q be $g(w)$ and let $g(w/n)$ be the density function of W_q subject to the condition that there are n customers in the system or there are (n -1) customers in the queue apart from one customer receiving service.

Now $g(w/n)$ = pdf of sum of n service times [one residual service time + (n -1) full service times]

$$= \frac{\mu^n}{(n-1)!}\, e^{-\mu w}\, w^{n-1} : w > 0$$

Therefore $g(w) = \sum_{n=1}^{\infty} \frac{\mu^n}{(n-1)!}\, e^{-\mu w}\, w^{n-1} \left(\frac{\lambda}{\mu}\right)^n \left(1 - \frac{\lambda}{\mu}\right)$

$$= \lambda \left(1 - \frac{\lambda}{\mu}\right) e^{-\mu w} \sum_{n=1}^{\infty} \frac{1}{(n-1)!} (\lambda w)^{n-1}$$

$$= \frac{\lambda}{\mu} (\mu - \lambda)\, e^{-\mu w}\, e^{-\lambda w}$$

NOTES

$$= \frac{\lambda}{\mu}(\mu - \lambda)\, e^{-(\mu-\lambda)w} : w > 0 \qquad (9)$$

and $g(w) = 1 - \frac{\lambda}{\mu}$, when $w = 0$

9.Average waiting time of a customer in the queue

$$E(W_q) = \frac{\lambda}{\mu}(\mu - \lambda) \int_0^\infty w\, e^{-(\mu-\lambda)w}\, dw$$

$$= \frac{\lambda}{\mu} \int_0^\infty x\, e^{-x} \frac{dx}{(\mu - \lambda)}.$$

$$= \frac{\lambda}{\mu(\mu - \lambda)} [x(-e^{-x}) - e^{-x}]_0^\infty$$

$$= \frac{\lambda}{\mu(\mu - \lambda)} \qquad (10)$$

10. Average waiting time of a customer in the queue, if he has to wait

$$E(W_q / W_q > 0) = \frac{E(W_q)}{P(W_q > 0)}$$

$$= \frac{E(W_q)}{1 - P(W_q = 0)}$$

$$= \frac{E(W_q)}{1 - P(\text{no customer in the queue})}.$$

$$= \frac{E(W_q)}{1 - P_0}$$

$$= \frac{\lambda}{\mu(\mu - \lambda)} \times \frac{\mu}{\lambda} \quad (\text{since } P_0 = 1 - \frac{\mu}{\lambda})$$

$$= \frac{1}{(\mu - \lambda)}$$

Little's formula

1) $E(N_s) = L_s = \frac{\lambda}{(\mu - \lambda)} = \lambda E(W_s)$

NOTES

2) $E(N_q) = L_q = \frac{\lambda^2}{\mu(\mu - \lambda)} = \lambda E(W_q)$

3) $E(W_s) = E(W_q) + \frac{1}{\mu}$

4) $E(N_s) = E(N_q) + \frac{\lambda}{\mu}$

If any one of the equations $E(N_s)$, $E(N_q)$, $E(W_s)$ and $E(W_q)$ is known, the other three can be found out using the relations given above.

5.8 CHARACTERISTICS OF INFINITE CAPACITY , MULTIPLE SERVER POISSON QUEUE MODEL II (M/M/s) : (∞/FIFO), when $\lambda_n = \lambda$ and $\mu_n = \mu$ ($\lambda < s\mu$)

1.Values of P_0 and P_n

For the Poisson queue system P_n is given by

$$P_n = \frac{\lambda_0 \lambda_1 \lambda_2 \ldots \lambda_{n-1}}{\mu_1 \mu_2 \mu_3 \ldots \mu_n} \times P_0 \quad n \geq 1 \qquad (1)$$

$$P_0 = \left[1 + \sum_{n=1}^{\infty} \frac{\lambda_0 \lambda_1 \lambda_2 \ldots \lambda_{n-1}}{\mu_1 \mu_2 \mu_3 \ldots \mu_n} \right]^{-1} \qquad (2)$$

If there is a single server, $\mu_n = \mu$ for all n. But there are s servers working independently of each other,

If there be less than s customers, i.e., if $n < s$, only n of the servers will be busy and the others idle and hence the mean service rate will be $n\mu$.

If $n \geq s$, all the servers will be busy and hence the mean service rate = $s\mu$.

hence $\mu_n = n\mu$, if $0 \leq n < s$

$= s\mu$, if $n \geq s$ (3)

Using (3) an (1) and (2), we have

$$P_n = \frac{\lambda^n}{1\mu.2\mu.3\mu \ldots n\mu} \times P_0, \quad \text{if } 0 \leq n < s$$

$$= \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n \times P_0, \quad \text{if } 0 \leq n < s \qquad (4)$$

and $P_n = \frac{\lambda^n}{\{1\mu.2\mu.3\mu \ldots (s-1)\mu\}\{s\mu.s\mu.s\mu \ldots (n-s+1) \text{ times}\}} \times P_0$

NOTES

$$P_n = \frac{\lambda^n}{(s-1)!\,\mu^{s-1}(s\mu)^{n-s+1}} \times P_0$$

$$\boxed{\mathbf{P_n = \frac{1}{s!s^{n-s}}\left(\frac{\lambda}{\mu}\right)^n X P_0, \quad if \quad n \geq s}} \tag{5}$$

Now P_0 is given by $\sum_{n=0}^{\infty} P_n = 1$

$$\text{i.e.,} \left[\sum_{n=0}^{s-1} \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n + \sum_{n=0}^{\infty} \frac{1}{s!s^{n-s}}\left(\frac{\lambda}{\mu}\right)^n\right] \times P_0 = 1$$

$$\text{i.e.,} \left[\sum_{n=0}^{s-1} \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n + \frac{s^s}{s!}\sum_{n=s}^{\infty}\left(\frac{\lambda}{\mu s}\right)^n\right] \times P_0 = 1$$

$$\text{i.e.,} \left[\sum_{n=0}^{s-1} \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n + \frac{s^s}{s!}\left(\frac{\lambda}{\mu s}\right)^s \frac{1}{1-\frac{\lambda}{\mu s}}\right] \times P_0 = 1$$

$$\text{i.e.,} \left[\sum_{n=0}^{s-1} \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n + \frac{1}{s!\left(1-\frac{\lambda}{\mu s}\right)}\left(\frac{\lambda}{\mu}\right)^s\right] \times P_0 = 1$$

$$\boxed{\text{or } \mathbf{P_0 = \frac{1}{\left[\sum_{n=0}^{s-1} \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n\right] + \left[\frac{1}{s!\left(1-\frac{\lambda}{\mu s}\right)}\left(\frac{\lambda}{\mu}\right)^s\right]}}} \tag{6}$$

2. Average number L_q of customers in the queue or average length of the queue:

$$L_q = E(N - s) = \sum_{x=s}^{\infty} (n - s) \times P_n$$

$$= \sum_{x=0}^{\infty} x\, P_{x+s}$$

$$= \sum_{x=0}^{\infty} x \times \frac{1}{s!s^x}\left(\frac{\lambda}{\mu}\right)^{s+x} \times P_0$$

$$= \frac{1}{s!}\left(\frac{\lambda}{\mu}\right)^s P_0 \sum_{x=0}^{\infty} x \left(\frac{\lambda}{\mu s}\right)^x$$

NOTES

$$= \frac{1}{s!}\left(\frac{\lambda}{\mu}\right)^{s}\frac{\lambda}{\mu s}P_0\left(\frac{1}{1-\frac{\lambda}{\mu s}}\right)^{2}$$

$$= \frac{1}{s.\,s!}\left(\frac{\lambda}{\mu}\right)^{s+1}\left(\frac{1}{1-\frac{\lambda}{\mu s}}\right)^{2}P_0 \qquad (7)$$

3. Average number of customers in the system

By 4th Little's formula

$$E(N_s) = E(N_q) + \frac{\lambda}{\mu}$$

$$= \frac{1}{s.\,s!}\left(\frac{\lambda}{\mu}\right)^{s+1}\left(\frac{1}{1-\frac{\lambda}{\mu s}}\right)^{2}P_0 + \frac{\lambda}{\mu} \qquad (8)$$

This result can also be obtained by using the definition $E(N_s) = \sum_{n=0}^{\infty} n P_n$

4. Average time a customer has to spend in the system

By 1st Little's formula

$$E(N_s) = \lambda E(W_s)$$

$$\text{or } E(W_s) = \frac{E(N_s)}{\lambda}$$

$$= \frac{\frac{1}{s.\,s!}\left(\frac{\lambda}{\mu}\right)^{s+1}\left(\frac{1}{1-\frac{\lambda}{\mu s}}\right)^{2}P_0 + \frac{\lambda}{\mu}}{\lambda}.$$

$$= \frac{1}{\mu} + \frac{1}{\mu\, s.\,s!}\left(\frac{\lambda}{\mu}\right)^{s}\left(\frac{1}{1-\frac{\lambda}{\mu s}}\right)^{2}P_0 \qquad (9)$$

5. Average time a customer has to spend in the queue

By 2nd Little's formula

NOTES

$$E(N_q) = \lambda E(W_q)$$

or $$E(W_q) = \frac{E(N_q)}{\lambda}$$

$$= \frac{1}{s.\,s!} \frac{\left(\frac{\lambda}{\mu}\right)^{s+1}}{\left(1 - \frac{\lambda}{\mu s}\right)^2} P_0 .$$

$$= \frac{1}{\mu} \frac{1}{s.\,s!} \frac{\left(\frac{\lambda}{\mu}\right)^{s}}{\left(1 - \frac{\lambda}{\mu s}\right)^2} P_0 \qquad (10)$$

6. Probability that an arrival has to wait

Required probability = Probability that there are s or more customers in the system

$$P(W_s > 0) = P(N \geq s)$$

$$= \sum_{n=s}^{\infty} P_n = \sum_{n=s}^{\infty} \frac{1}{s!\, s^{n-s}} \left(\frac{\lambda}{\mu}\right)^n P_0$$

$$= \frac{1}{s!} \left(\frac{\lambda}{\mu}\right)^s P_0 \sum_{n=s}^{\infty} \left(\frac{\lambda}{\mu s}\right)^{n-s}$$

$$= \frac{\left(\frac{\lambda}{\mu}\right)^s P_0}{s! \left(1 - \frac{\lambda}{\mu s}\right)} \qquad (11)$$

7. Probability that an arrival enters the service without waiting.

Required probability = 1 – P(an arrival to wait)

$$= 1 - \frac{\left(\frac{\lambda}{\mu}\right)^s P_0}{s! \left(1 - \frac{\lambda}{\mu s}\right)} \qquad (12)$$

8. Mean waiting time in the queue for those who actually wait

$$E(W_q / W_s > 0) = \frac{E(W_q)}{P(W_s > 0)}$$

NOTES

Using (10) and (11)

$$= \frac{1}{\mu}\,\frac{1}{s.\,s!}\left(\frac{\lambda}{\mu}\right)^{s} P_0 \times \frac{s!\left(1-\frac{\lambda}{\mu s}\right)}{\left(\frac{\lambda}{\mu}\right)^{s} P_0}.$$

$$= \frac{1}{\mu s - \lambda} \qquad (13)$$

9. Probability that there will be someone waiting

Required probability = $P(N \overset{24}{12} s + 1)$

$$= \sum_{n=s+1}^{\infty} P_n = \sum_{n=s}^{\infty} P_n - P(N = s)$$

using (10) and (5)

$$= \frac{\left(\frac{\lambda}{\mu}\right)^{s} P_0}{s!\left(1-\frac{\lambda}{\mu s}\right)} - \frac{\left(\frac{\lambda}{\mu}\right)^{s}}{s!} P_0$$

$$= \frac{\left(\frac{\lambda}{\mu}\right)^{s} P_0 \left(\frac{\lambda}{\mu}\right)^{s}}{s!\left(1-\frac{\lambda}{\mu s}\right)}. \qquad (14)$$

10. Average number of customers (in non-empty queues), who have to actually wait

$L_w = E(N_q / N_q \overset{24}{12} 1)$

$= E(N_q) / P(N \overset{24}{12} s)$

$$= \frac{\left(\frac{\lambda}{\mu}\right)^{s+1} P_0\; s!\left(1-\frac{\lambda}{\mu s}\right)}{s.\,s!\left(1-\frac{\lambda}{\mu s}\right)^{2}\left(\frac{\lambda}{\mu}\right)^{s} P_0}.$$

$$= \frac{\frac{\lambda}{\mu s}}{1-\frac{\lambda}{\mu s}}. \qquad (15)$$

11) $P(W > t) = e^{-\mu t}\left(1 + \frac{(\lambda/\mu)^{s}[\,1 - e^{-\mu t(s-1-(\lambda/\mu))}]\,P_0}{s!\,(1-(\lambda/\mu s))\,(s-1-(\lambda/\mu))}\right)$

NOTES

5.9 CHARACTERISTICS OF FINITE CAPACITY , SINGLE SERVER POISSON QUEUE MODEL III [(M/M/1) : (k/FIFO)] model

1. Values of P_0 and P_n

For the Poisson queue system $P_n = P(N = n)$ in the steady state is given by the difference equations,

$\lambda_{n-1} P_{n-1} - (\lambda_n + \mu_n) P_n + \mu_{n+1} P_{n+1} = 0 : n > 0$

and $\lambda_0 P_0 + \mu_1 P_1 = 0: n = 0$

This model represents the situations in which the system can accommodate only a finite number k of arrivals. If a customer arrives and the queue is full, the customer leaves without joining the queue.

Therefore for this model,

$\mu_n = \mu, n = 1, 2,3, \ldots .$

and $\lambda_n = \begin{cases} \lambda, \text{ for } n = 0,1,2, \ldots (k-1) \\ 0, \text{ for } n = k, k+1, \ldots \end{cases}$

Using these values in the difference equations given above, we have

$\mu P_1 = \lambda P_0$ (1)

$\mu P_{n+1} = (\lambda + \mu) P_n - \lambda P_{n-1}$, for $1 \le n \le k-1$ (2)

$\mu P_k = \lambda P_{k-1}$, for $n = k$ (since P_{k+1} has no meaning) (3)

From (1), $P_1 = \frac{\lambda}{\mu} P_0$

from (2) $\mu P_2 = (\lambda + \mu) \frac{\lambda}{\mu} P_0 - \lambda P_0$

$P_2 = \left(\frac{\lambda}{\mu}\right)^2 P_0$ and so on

In general, $P_n = \left(\frac{\lambda}{\mu}\right)^n P_0$, for $0 \le n \le k-1$

From (3) $1 \le n \le k-1$

$P_k = \frac{\lambda}{\mu} \left(\frac{\lambda}{\mu}\right)^{k-1} P_0 = \left(\frac{\lambda}{\mu}\right)^k P_0$

Now $\sum_{n=0}^{k} P_n = 1$

NOTES

i.e., $P_0 \sum_{n=0}^{k} \left(\frac{\lambda}{\mu}\right)^n = 1$

i.e., $\frac{P_0 \left\{1 - \left(\frac{\lambda}{\mu}\right)^{k+1}\right\}}{1 - \frac{\lambda}{\mu}} = 1$

which is valid for $\lambda > \mu$

$$P_0 = \begin{cases} \dfrac{1 - \frac{\lambda}{\mu}}{1 - \left(\frac{\lambda}{\mu}\right)^{k+1}}, & \text{if } \lambda \neq \mu \quad (4) \\ \dfrac{1}{k+1}, & \text{if } \lambda = \mu \quad (5) \end{cases}$$

since $\lim_{\frac{\lambda}{\mu} \to 1} \dfrac{1 - \frac{\lambda}{\mu}}{1 - \left(\frac{\lambda}{\mu}\right)^{k+1}} = \dfrac{1}{k+1}$

$$P_n = \begin{cases} \left(\frac{\lambda}{\mu}\right)^n \dfrac{1 - \frac{\lambda}{\mu}}{1 - \left(\frac{\lambda}{\mu}\right)^{k+1}}, & \text{if } \lambda \neq \mu \quad (6) \\ \dfrac{1}{k+1}, & \text{if } \lambda = \mu \quad (7) \end{cases}$$

2. Average number of customers in the system

$$E(N) = \sum_{n=0}^{k} nP_n = \frac{1 - \frac{\lambda}{\mu}}{1 - \left(\frac{\lambda}{\mu}\right)^{k+1}} \sum_{n=0}^{k} n \left(\frac{\lambda}{\mu}\right)^n$$

$$= \frac{1 - \frac{\lambda}{\mu}}{1 - \left(\frac{\lambda}{\mu}\right)^{k+1}} \cdot \frac{\lambda}{\mu} \sum_{n=0}^{k} \frac{d}{dx}(x^n), \quad \text{where } x = \frac{\lambda}{\mu}$$

NOTES

$$= \frac{1 - \frac{\lambda}{\mu}}{1 - \left(\frac{\lambda}{\mu}\right)^{k+1}} \frac{d}{dx}\frac{(1 - x^{k+1})}{(1 - x)}$$

$$= \frac{(1 - x)}{1 - x^{k+1}} \frac{x}{} \frac{\{(1 - x)\{-(k + 1)x^k\} + (1 - x^{k+1})\}}{(1 - x)^2}$$

$$= \frac{x\,[1 - (k + 1)x^k + k\,x^{k+1}]}{(1 - x)\,(1 - x^{k+1})}$$

$$= \frac{x(1 - x^{k+1}) - (k + 1)(1 - x)\,x^{k+1}}{(1 - x)\,(1 - x^{k+1})}$$

$$= \frac{x}{1 - x} - \frac{(k + 1)\,x^{k+1}}{1 - x^{k+1}}$$

$$= \frac{\lambda}{\mu - \lambda} - \frac{(k + 1)\left(\frac{\lambda}{\mu}\right)^{k+1}}{1 - \left(\frac{\lambda}{\mu}\right)^{k+1}} \quad \text{if } \lambda \neq \mu \qquad (8)$$

and $E(N) = \sum_{n=0}^{k} \frac{n}{k+1} = \frac{k}{2}$ if $\lambda = \mu$ (9)

3. Average number of customers in the queue.

$$E(N_q) = E(N - 1) = \sum_{n=1}^{k} (n - 1)\,P_n$$

$$= \sum_{n=0}^{k} n\,P_n - \sum_{n=1}^{k} P_n$$

$$= E(N) - (1 - P_0) \qquad (10)$$

By 4th Little's formula

$$E(N_s) = E(N_q) + \frac{\lambda}{\mu}$$

$$\text{or } E(N_q) = E(N_s) - \frac{\lambda}{\mu}$$

which is true when the average arrival rate is λ throughout. Now we see that, in step (10),

$1 - P_0 \neq \frac{\lambda}{\mu}$, because the average arrival rate is λ as long as there is a vacancy in the queue and it is zero when the system is full.

NOTES

Hence we define the ***overall effective arrival rate***, denoted by ***λ' or λ_{eff}*** . by using step (10) and Little's formula as

$$\frac{\lambda'}{\mu} = 1 - P_0 \text{ or } \lambda' = \mu(1 - P_0) = \lambda(1 - P_n) \quad (11)$$

Thus we have (12)

$$E(N_q) = E(N_s) - \frac{\lambda'}{\mu}$$

which is the **modified Little's formula** for this model.

4. Average waiting times in the system and in the queue

By modified Little's formulas,

$$E(W_s) = \frac{1}{\lambda'} E(N_s) \quad (13)$$

$$\text{and } E(W_q) = \frac{1}{\lambda'} E(N_q) \quad (14)$$

where λ' is the effective arrival rate.

5.10 CHARACTERISTICS OF FINITE CAPACITY , SINGLE SERVER POISSON QUEUE MODEL IV [(M/M/s) : (k/FIFO)] model

1. Values of P_0 and P_n

$$P_n = \frac{\lambda_0 \lambda_1 \lambda_2 \ldots . \lambda_{n-1}}{\mu_1 \mu_2 \mu_3 \ldots . \mu_n} P_0 \text{ or } n \overset{24}{12} 1 \quad (1)$$

$$P_0 = 1 + \sum_{n=1}^{\infty} \frac{\lambda_0 \lambda_1 \lambda_2 \ldots . \lambda_{n-1}}{\mu_1 \mu_2 \mu_3 \ldots . \mu_n}^{-1} \quad (2)$$

For this [(M/M/s) : (k/FIFO)] model,

$$\lambda_n = \begin{cases} \lambda, \text{ for } n = 0, 1, 2, \quad k-1 \\ 0, \text{ for } n = k, k+1, \ldots . \end{cases}$$

$$\mu_n = \begin{cases} n\mu, \text{ for } n = 0, 1, 2, \ldots, s-1 \\ s\mu, \text{ for } n = s, s+1, \ldots . \end{cases}$$

using these values of λ_n and μ_n in (2) and noting that $1 < s < k$, we get

$$P_0^{-1} = \left\{ 1 + \frac{\lambda}{1!\mu} + \frac{\lambda^2}{2!\mu^2} + \ldots . + \frac{\lambda}{(s-1)!\mu^{s-1}} \right\} + \left\{ \frac{\lambda^s}{(s-1)!\mu^{s-1}\mu s} . \right.$$

NOTES

$$+ \frac{\lambda^{s+1}}{(s-1)!\mu^{s-1}(\mu s)^2} + \ldots\ldots + \frac{\lambda^{k}}{(s-1)!\mu^{s-1}(\mu s)^{k-s-1}} \Bigg\}$$

$$= \sum_{n=0}^{s-1} \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n + \frac{\lambda^s}{s!\,\mu^s}\left[1 + \frac{\lambda}{\mu s} + \left(\frac{\lambda}{\mu s}\right)^2 + \ldots\ldots + \left(\frac{\lambda}{\mu s}\right)^{k-s}\right]$$

$$P_0^{-1} = \sum_{n=0}^{s-1} \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n + \frac{1}{s!}\left(\frac{\lambda}{\mu}\right)^s \sum_{n=s}^{k}\left(\frac{\lambda}{\mu s}\right)^{n-s} \tag{3}$$

$$P_n = \begin{cases} \dfrac{1}{n!}\left(\dfrac{\lambda}{\mu}\right)^n P_0, & \text{for } n \le s \\ \dfrac{1}{s!\,s^{n-s}}\left(\dfrac{\lambda}{\mu}\right)^n P_0 & \text{for } s < n \le k \\ 0, & \text{for } n > k \end{cases} \tag{4}$$

2. Average queue length or average number of customers in the queue

$$E(N_q) = E(N - s) = \sum_{n=s}^{k} (n - s)\, P_n$$

$$= \frac{P_0}{s!} \sum_{n=s}^{k} (n - s) \frac{1}{s^{n-s}} \left(\frac{\lambda}{\mu}\right)^n \text{(using (4))}$$

$$= \frac{\left(\frac{\lambda}{\mu}\right)^s P_0}{s!} \sum_{x=0}^{k-s} x \left(\frac{\lambda}{\mu s}\right)^x$$

$$= \frac{\left(\frac{\lambda}{\mu}\right)^s P_0}{s!}\,\rho \sum_{x=0}^{k-s} x\,\rho^{x-1} \quad \text{where } \rho = \frac{\lambda}{\mu s}$$

$$= \frac{\lambda}{\mu}^{s} \frac{P_0}{s!}\,\rho \sum_{x=0}^{k-s} \frac{d}{d\rho}(\rho^x)$$

$$= \left(\frac{\lambda}{\mu}\right)^s \frac{P_0}{s!}\,\rho \frac{d}{d\rho}\left(\frac{1 - \rho^{k-s+1}}{1 - \rho}\right)$$

$$= \left(\frac{\lambda}{\mu}\right)^s \frac{P_0}{s!}\,\rho \left[\frac{-(1 - \rho)(k - s + 1)\rho^{k-s} + (1 - \rho^{k-s+1})}{(1 - \rho)^2}\right]$$

$$= \left(\frac{\lambda}{\mu}\right)^s \frac{P_0}{s!}\,\rho \left[\frac{(k - s)(1 - \rho)\rho^{k-s} - (1 - \rho)\rho^{k-s} + (1 - \rho^{k-s+1})}{(1 - \rho)^2}\right]$$

$$= P_0 \left(\frac{\lambda}{\mu}\right)^s \frac{\rho}{s!} \left[\frac{-(k - s)(1 - \rho)\rho^{k-s} + 1 - \rho^{k-s}(1 - \rho + \rho)}{(1 - \rho)^2}\right]$$

$$= P_0 \left(\frac{\lambda}{\mu}\right)^s \frac{\rho}{s!\,(1-\rho)^2} \left[1 - \rho^{k-s} - (k-s)(1-\rho)\rho^{k-s}\right] \quad (5)$$

where $\rho = \frac{\lambda}{\mu}$

3.Average number of customers in the system

$$E(N) = \sum_{n=0}^{k} n P_n$$

$$= \sum_{n=0}^{s-1} n P_n + \sum_{n=s}^{k} nP_n$$

$$= \sum_{n=0}^{s-1} n P_n + \sum_{n=s}^{k} (n-s)P_n + \sum_{n=s}^{k} sP_n$$

$$= \sum_{n=0}^{s-1} n P_n + E(N_q) + s \sum_{n=0}^{k} P_n - \sum_{n=s}^{s-1} P_n$$

$$= E(N_q) + s - \sum_{n=0}^{s-1} (s-n) P_n \quad \text{since } \sum_{n=0}^{k} P_n = 1 \quad (6)$$

Obviously $s - \sum_{n=0}^{s-1} (s-n) P_n \neq \frac{\lambda}{\mu}$, so that step (6) represents Little's formula.

In order to make (6) to assume the form of Little's formula, we define the overall effective arrival rate λ' or λ_{eff} as follows:

$$\frac{\lambda'}{\mu} = s - \sum_{n=0}^{s-1} (s-n) P_n .$$

$$\lambda' = \mu \left[s - \sum_{n=0}^{s-1} (s-n) P_n\right] \quad (7)$$

with this definition of λ', step (6) becomes

$$E(N) = E(N_q) + \frac{\lambda'}{\mu} \quad (8)$$

which is the **modified Little's formula** for this model

NOTES

NOTES

4. Average waiting time in the system and in the queue

By modified Little's formula

$$E(W_s) = \frac{1}{\lambda'} E(N) \quad (9)$$

$$\text{and } E(W_q) = \frac{1}{\lambda'} E(N_q) \quad (10)$$

Example 1: In a railway marshalling yard, goods trains arrive at a rate of 30 trains per day. Assuming that the inter-arrival time follows an exponential distribution and the service time (the time taken to hump a train) distribution is also exponential with a average of 36 minutes. Calculate

i) Expected queue size (line length)
ii) Probability that the queue size exceeds 10.

If the input of trains increases to an average of 33 per day, what will be the change in iI and ii) ?

Solution: (Model I)

Arrival rate $\lambda = 30$ trains per day

$= \frac{30}{60 \times 24} = 1/48$ trains per minute

Average of service time is given, that means $1/\mu$ is given

$1/\mu = 36$ minutes

Therefore $\mu = 1/36$ trains per minute.

i) Expected queue size (line length) $= L_S = \frac{\lambda}{\lambda - \mu} = 3$ trains.(omit the negative sign)

ii) Probability that the queue exceeds 10, $P(N > 10) = (\lambda/\mu)^{k+1} = (36/48)^{11} = 0.042$

Now if the input increase to 33 trains per day, then we have

$\lambda = \frac{33}{60 \times 24} = 11/480$ trains per minute and $\mu = 1/36$ trains per minute.

i) Expected queue size (line length) $= L_S = \frac{\lambda}{\lambda - \mu} = 4.71\frac{5}{8}$ 5 trains.

ii) Probability that the queue exceeds 10,

$P(N > 10) = (\lambda/\mu)^{k+1} = ((11 \times 36) / 480)^{11} = 0.121$

NOTES

Example 2: Arrivals at a telephone booth are considered to be Poisson with an average time of 12 minutes between one arrival and the next. The length of a phone call assumed to be distributed exponentially with mean 4 min.

a) Find the average number of persons waiting in the system.
b) What is the probability that a person arriving at the booth will have to wait in the queue?
c) What is the probability that it will take him more than 10 minutes altogether to wait for the phone and complete his call?
d) Estimate the fraction of the day when the phone will be in use?
e) The telephone department will install a second booth, when convinced that an arrival has to wait on the average for at the least 3 minutes for phone . By how much flowof arrivals should increase in order to justify a second booth?
f) What is the average length of the queue that forms time to time?

Solution: (Model I)

Mean inter arrival time = $1/\lambda = 12$ minutes
Therefore mean arrival rate, $\lambda = 1/12$ per minute
Mean service time $1/\mu = 4$ minutes
Therefore mean service rate $\mu = ¼$ per minute

a) $E(N) = L_S = \frac{\lambda}{\lambda - \mu} = 0.5$ customer

b) $P(W > 0) = 1 - P(W = 0)$
$= 1 - P(\text{no customer in the system})$
$= 1 - P_0$
$= 1 - [1 - (\lambda/\mu)] = \lambda/\mu = 1/3$

c) $P(W > 10) = e^{-(\mu - \lambda)\times 10} = e^{-(1/4 - 1/12)\times 10} = 0.1889$

d)P(the phone will be idle) = $P(N = 0) = P_0 = 1 - (\lambda/\mu) = 2/3$.

e)The second phone will be installed, if $E(W_q) > 3$

i.e., if $= \frac{\lambda}{\mu(\mu - \lambda)} > 3$

i.e., if $= \frac{\lambda_R}{\mu(\mu - \lambda_R)} > 3$ where λ_R is the required arrival rate.

i.e., if $= \frac{\lambda_R}{1/4(1/4 - \lambda_R)} > 3$

NOTES

i.e., if $\lambda_R > \frac{3}{4}(1/4 - \lambda_R)$

i.e., if $\lambda_R > 3/16 - 3/4\ \lambda_R$

i.e., if $\lambda_R > 3/28$

Hence the arrival rate should increase by 3/28 – 1/12 = 1/42 per minute, to justify a second phone.

f) Average length of a non – empty queue $L_w = \frac{\mu}{\mu - \lambda} = 1.5$ persons

Example 3: A petrol pump station has 4 pumps. The service times follow the exponential distribution with a mean of 6 minutes and cars arrive for service in a Poisson process at the rate of 30 cars per hour.

a) What is the probability that an arrival would have to wait in line?
b) Find the average waiting time, average time spent in the system and the average number of cars in the system?
c) For what percentage of time would a pump be idle on an average?

Solution: (Model II)

$s = 4, \lambda = 30$/hour
$1/\mu = 6$minutes $= 6/60$ hours $= 1/10$ hours
$\mu = 10$ / hour

a) P(an arrival has to wait) = P(W > 0)

$$= \frac{(\lambda/\mu)^s P_0}{s!(1 - (\lambda/\mu s))}$$

$$= \frac{3^4 \times P_0}{24 \times (1 - \frac{3}{4})} = 13.5 \times P_0$$

$$\text{Now } P_0 = \left(\left\{ \sum_{n=0}^{S-1} \frac{1}{n!} (\lambda/\mu)^n \right\} + \frac{(\lambda/\mu)^s}{s!(1 - (\lambda/\mu s))} \right)^{-1}$$

$$= \left((1 + 3 + \tfrac{1}{2} \times 9 + 1/6 \times 27) + \frac{3^4}{24 \times (1 - 1/3)} \right)^{-1}$$

$$= 0.0377$$

NOTES

Therefore $P(W > 0) = 13.5 \times P_0 = 13.5 \times 0.0377 = 0.5090$

b) $E(W_q) = \frac{1}{\mu} \quad \frac{1}{s \times s!} \times \frac{(\lambda/\mu)^s}{(1 - (\lambda/\mu s))^2} \times P_0$

$$= \frac{1}{10 \times 4 \times 24} \times \frac{3^4}{(1 - 3/4)^2} \times 0.0377$$

$= 0.0509$ hours

$E(W_s) = E(W_q) + 1/\mu = 0.0509$ hours ++ 1/10 hours

$= 0.1509$

$L_s = \lambda E(W_s) = 30 \times 0.1509 = 4.527$ cars.

c)The fraction of time when the pumps are busy
= traffic intensity
$= (\lambda/\mu s)$
$= 30/(10 \times 4)$
$= ¾$

Therefore the fraction of time when the pumps are idle = ¼

Therefore, required percentage = 25%

Example 4: There are three typists in an office. Each typist can type an average of 6 letters per hour. If the letters arrive for being typed at the rate of 15 letters per hour,

a) What fraction of the time all the typists will be busy?
b) What is the average number of letters waiting to be typed?
c) What is the average time a letter has to spend for waiting and for being typed?
d) What is the probability that a letter will take longer than 20 minutes waiting to be typed and being typed?

Solution: (Model II)

$s = 3$, $\lambda = 15$/hour
$\mu = 6$/ hour

a)P(all the typists are busy) = P(there are 3 or more customers in the system)
$= P(N \overset{24}{12} 3)$

NOTES

$$= \frac{(\lambda/\mu)^s P_0}{s!(1 - (\lambda/\mu s))}$$

$$= \frac{(15/6)^3 P_0}{6 \times (1 - (15/6 \times 3))}.$$

$$\text{Now } P_0 = \left\{ \left\{ \sum_{n=0}^{S-1} \frac{1}{n!} (\lambda/\mu)^n \right\} + \frac{(\lambda/\mu)^s}{s!(1 - (\lambda/\mu s))} \right\}^{-1}$$

$$= \left((1 + 2.5 + \tfrac{1}{2} \times (2.5)^2) + \frac{(2.5)^3}{6 \times (1 - 5/6)} \right)^{-1}$$

$$= 0.0449$$

$$\text{Therefore } P(N \overset{24}{12} 3) = \frac{(15/6)^3 P_0}{6 \times (1 - (15/6 \times 3))}$$

$$= \frac{(15/6)^3 \times 0.0449}{6 \times (1 - (15/6 \times 3))}$$

$$= 0.7016$$

$$\text{b) } E(W_q) = \frac{1}{s \times s!} \times \frac{(\lambda/\mu)^{s+1}}{(1 - (\lambda/\mu s))^2} \times P_0$$

$$= \frac{1}{3 \times 6} \times \frac{(2.5)^4}{(1 - 2.5/3)^2} \times 0.0449$$

$$= 3.5078$$

$$\text{c) } E(W_s) = \frac{E(N_s)}{\lambda} \quad \text{by Little's formula}$$

$$= \frac{1}{\lambda} E(N_q) + \frac{\lambda}{\mu} \quad \text{again by Little's formula}$$

$$= \frac{1}{15} \{3.5078 + 2.5\}$$

$$= 0.4005 \text{ hours}$$

$$\text{d) } P(W > t) = e^{-\mu t} \left\{ 1 + \frac{(\lambda/\mu)^s [1 - e^{-\mu t(s - 1 - (\lambda/\mu))}] P_0}{s!(1 - (\lambda/\mu s))(s - 1 - (\lambda/\mu))} \right\}$$

NOTES

$$P(W > 1/3) = e^{-6 \times 1/3}\left[1 + \frac{(2.5)^3[1 - e^{-(-2 \times -0.5)}] \times 0.0449}{6\,(1 - (2.5/3))\,(-0.5)}\right]$$

$$= e^{-2}\left[1 + \frac{0.7016\,(1 - e)}{(-0.5)}\right]$$

$$= 0.4616$$

Example 5: The local one-person barber shop can accommodate a maximum of 5 people at a time (4 waiting and 1 getting hair cut). Customers arrive according to a Poisson distribution with mean 5 per hour. The barber cuts hair at an average rate of 4 per hour.

a) What percentage of time is barber idle?
b) What fraction of the potentials customers are turned away?
c) What is the expected umber of customers waiting for a hair-cut?
d) How much time can a customer expect to spend in the barber shop?

Solution: (Model III, $\lambda \frac{4}{4} \mu$)

$\lambda = 5$, $\mu = 4$, $k = 5$

a) P(the barber is idle) = P(N =0)

$$= P_0$$

$$= \frac{1 - (\lambda/\mu)}{1 - (\lambda/\mu)^{k+1}}$$

$$= \frac{1 - 5/4}{1 - (5/4)^6}$$

$$= 0.0888$$

Percentage of time when the barber is idle = $8.88 \frac{5}{8} 9$

b)P(a customer is turned away) = P(N > 5)

$$= (\lambda/\mu)^n \frac{1 - (\lambda/\mu)}{1 - (\lambda/\mu)^{k+1}}$$

$$= (5/4)^5 \frac{1 - (5/4)}{1 - (5/4)^6}$$

$$= \frac{3125}{11529} = 0.2711$$

Therefore, 0.2711potential customers are turned away.

NOTES

c) $E(N_q) = E(N) - (1 - P_0)$

$$= \frac{\lambda}{\mu - \lambda} - \frac{(k+1)(\lambda/\mu)^{k+1}}{1 - (\lambda/\mu)^{k+1}} - (1 - P_0)$$

$$= \frac{\lambda}{\mu - \lambda} - \frac{6 \times (5/4)^6}{1 - (5/4)^6} - (1 - 0.0888)$$

$$= \frac{6 \times (15625/4096)}{11529/4096} - 5.9112 = 2.2 \text{ customers.}$$

d) $E(W) = \frac{1}{\lambda'} E(N)$

where $\lambda' = \mu(1 - P_0)$

$$E(W) = \frac{1}{\mu(1 - P_0)} \times E(N) = 3.1317 / 3.6448 = 0.8592 \text{ hours}$$

Example 6: At a railway station, only one train is handled at a time. The railway yard is sufficient only fro two trains to wait, while the other is given signal to leave the station. Trains arrive at the station at an average rate of 6 per hour and the railway station can handle them on an average of 6 per hour. Assuming Poisson arrivals and exponential Poisson distribution, find the probabilities fro the numbers of trains in the system. Also find the average waiting time of a new train coming into the yard. If the handling rate is doubled, how will the above results get modified?

Solution:(Model III)

i) $\lambda = 6$ per hour, $\mu = 6$ per hour, $k = 2 + 1 = 3$

Here $\boldsymbol{\lambda = \mu}$, $P_0 = 1 / k + 1 = ¼$

$P_n = 1 / k + 1 = ¼$ for n = 1,2,3

$E(N) = k/2 = 1.5$

$E(W) = \frac{1}{\lambda'} E(N)$

where $\lambda' = \mu(1 - P_0)$

$$E(W) = \frac{1}{\mu(1 - P_0)} \times E(N)$$

$$= \frac{1}{6 \times (3/4)} \times 1.5 = 0.3333 \text{ hours} = 20 \text{ minutes}$$

ii)if the handling rate is doubled,

NOTES

$\lambda = 6$ per hour, $\mu = 12$ per hour, $k = 3$

$\lambda \frac{4}{4} \mu$

$$P_0 = \frac{1 - (\lambda/\mu)}{1 - (\lambda/\mu)^{k+1}}$$

$$= \frac{1 - ½}{1 - (1/2)^4}$$

$$= 0.5333$$

$$P_n = (\lambda/\mu)^n \frac{1 - (\lambda/\mu)}{1 - (\lambda/\mu)^{k+1}}$$

$$= (8 / 15)(1/2)^n, \text{ for } n = 1,2,3.$$

$$E(N) = \frac{\lambda}{\mu - \lambda} - \frac{(k+1)(\lambda/\mu)^{k+1}}{1 - (\lambda/\mu)^{k+1}}$$

$$= 1 - \frac{4 \times (1/2)^4}{1 - (1/2)^4} = 1 - (4/15) = 0.7333 \text{ trains}$$

$$E(W) = \frac{1}{\lambda'} E(N)$$

where $\lambda' = \mu(1 - P_0)$

$$E(W) = \frac{1}{\mu(1 - P_0)} \times E(N)$$

$$= \frac{1}{12(1 - (8/15))} \times (0.7333) = 0.13095 \text{ hours}$$

Example 7: A 2-person barber shop has 5 chairs to accommodate waiting customers. Potential customers who arrive when all 5 chairs are full, leave without entering barber shop. Customers arrive at the average rate of 4 per hour and spend an average of 12 minutes in the barber's chair. Compute P_0, P_1, P_7, E(Nq) and E(W).

Solution: (Model IV)

$\lambda = 4$ per hour, $\mu = 5$ per hour, $s = 2$, $k = 2 + 5 = 7$

NOTES

a) $P_0 = \left[\left\{ \sum_{n=0}^{S-1} \frac{1}{n!} (\lambda/\mu)^n \right\} + \frac{(\lambda/\mu)^s}{s!} \sum_{n=s}^{k} (\lambda/\mu s)^{n-s} \right]^{-1}$

$= \sum_{n=0}^{1} \left[\frac{1}{n!} (4/5)^n + (1/2)(4/5)^2 \sum_{n=2}^{7} (2/5)^{n-2} \right]^{-1}$

$= 1 + 4/5 + 8/25\{1 + 2/5 + (2/5)^2 + (2/5)^3 + (2/5)^4 + (2/5)^5\}^{-1}$

$= [\, 9/5 + 8/25\{ 1 - (0.4)^7/1 - 0.4 \}]^{-1} = 0.4287$

b) $P_n = \frac{1}{n!} \left(\frac{\lambda}{\mu}\right)^n P_0$, for $n \leq s$

Therefore $P_1 = (4/5) \times 0.4287 = 0.3430$

c) $P_n = \frac{1}{s!.s^{n-s}} \left(\frac{\lambda}{\mu}\right)^n P_0$, for $s < n \leq k$

$P_7 = \frac{1}{2 \times 2^{7-2}} \times (4/5)^7 \times 0.4287$

$= 0.0014$

d) $E(N_q) = P_0 (\lambda/\mu)^s \frac{\rho}{s!\,(1-\rho)^2} [1 - \rho^{k-s} - (k-s)(1-\rho)\rho^{k-s}]$ (where $\rho = \lambda/\mu s$)

$= (0.4287).(0.8)^2 \frac{(0.4)}{2 \times (0.6)^2} [1 - (0.4)^5 - 5 \times 0.6 \times (0.4)^5]$

$= 0.15$ customers

e) $E(N) = E(N_q) + s - \sum_{n=0}^{S-1} (s-n)P_n$

$= 0.1462 + 2 - \sum_{n=0}^{1} (2-n)P_n$

$= 2.1462 - (2 \times P_0 + 1 \times P_1)$

$= 2.1462 - (2 \times 0.4287 + 1 \times 0.3430)$

$= 0.95$ customers

NOTES

$E(W) = \frac{1}{\lambda'} E(N)$

where $\lambda' = \mu(s - \sum_{n=0}^{s-1} (s-n)P_n)$

$= 4[2 - (2 \times 0.4287 + 1 \times 0.3430)]$

$= 3.1984$

Therefore E(W) = 0.9458/3.1984 = 0.2957 hours

Example 8 : A car servicing station has 2 bays where service can be offered simultaneously. Because of space limitation, only 4 cars are accepted for servicing. The arrival pattern is Poisson with 12 cars per day. The service time in both the bays is exponentially distributed with $\mu = 8$ per bay. Find the average number of cars in the service station, the average number of cars waiting for service and the average time a car spends in the system.

Solution: (Model IV)

$\lambda = 12$ per day, $\mu = 8$ per day, $s = 2$, $k = 4$

a) $P_0 = \left(\left\{ \sum_{n=0}^{S-1} \frac{1}{n!} (\lambda/\mu)^n \right\} + \frac{(\lambda/\mu)^s}{s!} \sum_{n=s}^{k} (\lambda/\mu s)^{n-s} \right)^{-1}$

$= 1 + 1.5/1 + \frac{1}{2} \times (1.5)^2 \{1 + (.75) + (.75)^2\}^{-1}$

$= 0.1960$

$E(N_q)$ = Average number of cars waiting for service

$= P_0 (\lambda/\mu)^s \frac{\rho}{s!(1-\rho)^2} [1 - \rho^{k-s} - (k-s)(1-\rho)\rho^{k-s}]$ (where $\rho = \lambda/\mu s$)

$= (0.1960) \times (1.5)^2 \times \frac{(0.75)}{2 \times (0.25)^2} [1 - (0.75)^5 - 2 \times 0.25 \times (0.75)^5]$

$= 0.4134$ car

b) E(N) = Average number of cars in the system

$= E(N_q) + s - \sum_{n=0}^{S-1} (s-n)P_n$

$= 0.4134 + 2 - \sum_{n=0}^{1} (2-n)P_n$

NOTES

$= 2.4134 – (2\ P_0 + P_1)$

$= 2.4134 – (2 \times 0.1960 + 1.5 \times 0.1960)$

$= 1.73$ cars

c) $E(W_s) = \frac{1}{\lambda'} E(N_s)$

where $\lambda' = \mu(s – \sum_{n=0}^{s-1} (s-n)P_n)$

$= 8[2 – (2\ P_0 + P_1)]$

$= 10.512$

$E(W_s) = 1.73/10.512 = 0.1646$ day

Example 9: In a given M/M/1 queueing system the average arrivals is 4 customers per minute : $\rho = 0.7$. What are 1) mean number of customers L_s in the system 2) mean number of customers L_q in the queue 3) probability that the server is idle 4) mean waiting time W_s in the system.

Solution:

Given $\rho = 0.7$ $(\rho = \lambda/\mu)$

1) $L_s = \frac{\rho}{1-\rho}. = \frac{0.7}{1-0.7}. = 2.333$

2) $L_q = \frac{\rho^2}{1-\rho} = 1.6333$

3) $P_0 = 1 – \rho = 1 – 0.7 = 0.3$

4) $W_s = \frac{L_s}{\lambda} = 2.333/4 = 0.5833$

Example 10: Customers arrive at a watch repair shop according to a Poisson process at a rate of one per every 10 minutes, and the service time is an exponential random variable with mean 8 minutes.

1) Find the average number of customers L_s in the shop.
2) Find the average time a customer spends in the shop W_s
3) Find the average number of customers in the queue L_q
4) What is the probability that the service is idle.

NOTES

Solution: (Model I)

Arrival rate $\lambda = 1$ per every 10 minutes
Therefore $\lambda = 1/10$

Average of service time is given, that means $1/\mu$ is given
$1/\mu = 8$ minutes
$\mu = 1/8$

1) $L_S = \frac{\lambda}{\lambda - \mu} = 4$ customer

2) $W_s = \frac{1}{\lambda - \mu} = 40$ minutes

3) $L_q = L_s - (\lambda / \mu) = 4 - 8/10 = 3.2 \tilde{} 3$ customers.

4) $P_0 = 1 - (\lambda / \mu) = 1 - (4/5) = 0.2$

Example 11: A duplicating machine maintained for office use is operated by an office assistant. If the jobs arrive at a rate of 5 per hour and the time to complete each job varies according to an exponential distribution with mean 6 minutes, find the percentage of idle time of the machine in a day. Assume that jobs arrive according to a Poisson process.

Solution:

$\lambda = 5$ per hour
$1/\mu = 6$ minutes $= 6/60$ hours
$\mu = 10$ per hour
P(machine is idle) $= P(N = 0) = 1 - (\lambda / \mu) = ½$

Therefore the percentage of idle time = 50%

Example 12: What is the probability that an arrival to an infinite capacity 3 server Poisson process queue with $\lambda / (c\mu) = 2/3$ and $P_0 = 1/9$ enters the service without waiting?

Solution:

P(Without waiting) $= 1 - P(W > 0)$

$$= 1 - \frac{(\lambda/\mu)^s P_0}{s!(1 - (\lambda/\mu s))}$$

$$= \frac{2^3 \times 1/9}{6 \times (1 - 2/3)} = 0.5556$$

NOTES

Example 13: In a given M/M/1/ ∞ /FCFS queue, $\rho = 0.6$, what is the probability that the queue contains 5 or more customers?

Solution:

$P(X = 5) = (\rho)^5 = (0.6)^5 = 0.0467$

Example 14: What is the effective arrival rate of M/M/1/4/FCFS queueing model when

$\lambda = 2$ and $\mu = 5$.

Solution:

Overall effective arrival rate $= \lambda_{eff} = \lambda(1 - P_n)$

$$= \left(\lambda \; 1 - (\lambda/\mu)^n \; \frac{1 - (\lambda/\mu)}{1 - (\lambda/\mu)^{k+1}} \right)$$

$$\lambda_{eff} = \lambda(1 - P_4)$$

$$= \left(\lambda \; 1 - (\lambda/\mu)^4 \; \frac{1 - (\lambda/\mu)}{1 - (\lambda/\mu)^5} \right)$$

$$= 2 \; [1 - 2^4 \; (1\text{-}2 \; / 1 \; \text{-}2^5)]$$

$$= 0.9677$$

Example 15: Consider an M/M/1 queueing system. If $\lambda = 6$ and $\mu = 8$, find the probability of at least 10 customers in the system.

Solution:

$$P(X = 10) = \sum_{n=10}^{\infty} P_n = \sum (1 - 6/8)(6/8)^n = (6/8)^{10} = (3/4)^{10}$$

Example 16: A bank has two tellers working on saving accounts. The first teller handles withdrawals only. The second teller handles deposits only. It has been found that the service time distributions for both deposits and withdrawals are exponential with mean service time of 3 minute per customer. Depositors are found to arrive in a Poisson fashion throughout the day with mean arrival rate of 16 per hour. Withdrawers also arrive in a Poisson fashion with mean arrival rate of 14 per hour. What would be the effect on the average waiting time for the customers if each teller could handle both withdrawals and deposits. What would be the effect, if this could only be accomplished by increasing the service time to 3.5 minutes?

Solution:

NOTES

When there is a separate channel for the depositors, $\lambda_1 = 16$/hour, $\mu = 20$/hour
Therefore $E(W_q$ for depositors$) = \frac{\lambda_1}{\mu(\mu - \lambda_1)}$ (Model I)

$$= \frac{16}{20\,(20 - 16)} = 1/5 \text{ hours or } 12 \text{ minutes}$$

When there is a separate Poisson channel for the withdrawers, $\lambda_2 = 14$/hour, $\mu = 20$/ hour Therefore E(W_q for withdrawers) $= \frac{\lambda_2}{\mu(\mu - \lambda_2)}$ (Model I)

$$= \frac{14}{20\,(20 - 14)} = 7/60 \text{ hours or } 7 \text{ minutes}$$

If both tellers do both service, (Model II)

$s = 2$, $\mu = 20$/hour, $\lambda = \lambda_1 + \lambda_2 = 30$ /hour

$$E(W_q) = \frac{1}{\mu} \quad \frac{1}{s \times s!} \times \frac{(\lambda/\mu)^s}{(1 - (\lambda/\mu s))^2} \times P_0$$

$$= \frac{1}{20 \times 2 \times 2} \times \frac{(1.5)^2}{(1 - 7.5)^2} \times P_0$$

$$= 0.45 \times P_0$$

$$\text{Now } P_0 = \left(\left\{ \sum_{n=0}^{S-1} \frac{1}{n!} (\lambda/\mu)^n \right\} + \frac{(\lambda/\mu)^s}{s!(1 - (\lambda/\mu s))} \right)^{-1}$$

$$= \left((1 + 1.5 + \frac{(2.5)^2}{2 \times 0.25}) \right)^{-1}$$

$$= 1/7$$

Therefore $E(W_q) = 0.45 \times P_0 = 0.45 \times 1/7 = 0.0643$ hours

Hence if both tellers do both types of service, the customers get benefited as their waiting time is considerably reduced.

Now if both tellers do both types of service but with increased service time, $s = 2$, $\lambda = 30$, $\mu = 60/3.5 = 120/7$ per hour

NOTES

$$E(W_q \text{ of any customer}) = \frac{7}{120} \times \frac{1}{2 \times 2} \times \frac{(1.75)^2}{(1 - 7/8)^2} \times P_0 = 2.86\, P_0$$

$$P_0 = \left[1 + 1.75 + \frac{(1.75)^2}{2 \times 1/8}\right]^{-1} = 1/15$$

$E(W_q \text{ of any customer}) = 2.86\, P_0 = 2.86 \times 1/15 = .1907$ hours

If this arrangement is adopted, withdrawers stand to lose as their waiting time is increased considerably and depositors get slightly benefited.

How you understood ?

1. What are the characteristics of a queueing theory?
2. What do the letters in the symbolic representation (a/b/c):(d/e) of a queueing model represent?
3. What do you mean by transient state and steady state?
4. Write down the difference formulas for P_0 and P_n in a Poisson queue system in the steady-state.
5. Write down the Little's formulas that hold good for all the Poisson queue models.
6. Write down the formula for P_n in terms of P_0 for the (M/M/s):(8/FIFO) queueing System.
7. How are Ns and Nq related in an (M/M/1):(k/FIFO) queueing system.
8. Define effective arrival rate with respect to (M/M/s):(8/FIFO) queueing System.

TRY YOURSELF !

1) Customers arrive in a one- man barber shop according to a Poisson process with a mean interval time of 12 minutes. Customers spend an average of 10 minutes in the barber's chair.
   a) what is the expected number of customers in the barber shop an din the queue?
   b) Calculate the percentage of times an arrival can walk straight into the barber's chair without having to wait?
   c) How much time can a customer expect to spend in the barber's shop?
   d) Management will provide another chair and hire another barber, when a customer's waiting time in the shop exceeds 1.25 hours. How much the average rate of arrivals increase to warrant a second barber?
   e) What is the average time customers spend in the queue?
   f) What is the probability that the waiting time in the system is greater than 30 minutes?
   g) Calculate the percentage of customers who have to wait prior to getting into the barber's chair?
   h) What is the probability that more than 3 customers are in the system?

(Solution: 4.17, 1/6, 1/300 per minute, 50minute, 0.6065, 83.33, 0.4823)

NOTES

2) At what average rate must a clerk in a super market work in order to ensure a probability of 0.90 that the customer will not wait longer than 12 minute? It is assumed that there is only one counter at which customers arrive in a Poisson fashion at an average rate of 15 per hour and that the length of the service by the clerk has an exponential distribution.

(Solution: 24 customers per hour)

3) If the people arrive to purchase cinema tickets at the average rate of 6 perminute, it takes an average of 7.5 seconds to purchase a ticket. If a person arrives 2 minutes before the picture starts and it takes exactly 1.5 minute to reach the correct seat after purchasing the ticket,

a) Can he expect to be seated from the starting of the picture?
b) What is the probability that he will be seated from the starting of the picture?
c) How early must he arrive in order to be 99% sure of being seated for the start of the picture?

(Solution: 2 minutes, 0.63, 2.65 minutes)

4) Given an average arrival rate of 20 per hour, is it better for a customer to get service at a single channel with mean service rate of 22 customers per hour or at one of two channels in parallel with mean service rate of 11 customers per hour for each of the two channels. Assume both queues to be Poisson type.

5) A telephone company is planning to install telephone booths in a new airport. It has established the policy that a person should not have to wait more than 10% of the times he tries t use a phone. The demand for use is estimated be Poisson with an average of 0 per hour. The average phone call have an exponential distribution with a mean time of 5 minutes. How many phone booths should be installed?

(Solution: 6 booths should be installed)

6) A supermarket has two girls attending the sales at the counters. If the service time for each customer is exponential with mean 4 minutes and if people arrive in Poisson fashion at the rate of 10 per hour,
a) what is the probability that the customer has to wait for the service
b) what is the expected percentage of idle time for each girl.
c) if the customer has to wait in the queue, what is the expected length of his waiting time?

(Solution: 1/6, 67, 3 minutes)

7) Patients arrive at a clinic according to a Poisson distribution at a rate of 30 patients per hour. The waiting room does not accommodate more than 14 patients. Examination time per patient is exponential with mean rate of 20 per hour
a) Find the effective arrival rate at the clinic.
b) What is the probability that an arriving patient will not wait?

NOTES

c) What is the expected waiting time until a patient is discharged from the clinic?
(Solution: 19.98 per hour, 0.00076, 0.65 hours or 39 minutes)

8) A group of engineers has 2 terminals available to aid in their calculations. The average computing job requires 20 minutes of terminal time and each engineer requires some computation about once every half an hour. Assume that these are distributed according to an exponential distribution. IF there are 6 engineers in the group, find
a) the expected number of engineers waiting to use one of the terminals and in the computing centre and
b) the total time lost per day.
(Solution: 0.75, $16 \times 0.0398 = 0.6368$ hours)

REFERENCES:

1. T.Veerarajan, "Probability, statistics and Random Process ", Tata McGraw Hill, 2002.
2. P.Kandasamy, K. Thilagavathi and K. Gunavathi,"Probability, Random Variables and Random processors", S. Chand, 2003.
3. S.C Gupta and V.K Kapoor,"Fundementals of Mathemetical Statistics", Sultan Chand & Sons, 2002

SUMMARY:

At the end of this course you will be having fundenental knowledge of the basic probability concepts.You will be having well-founded knowledge of standard distributions which can describe real life phenomena. You would have acquiredskills in handling situations in involving more than one random and functions of random variable. You are exposed to basic characteristic features of a queueing system and would have acquired skills in analyzing queueing models.You are provided with necessary mathematical support and confidence to tackle real life problems.