NETAJI SUBHAS OPEN UNIVERSITY

STUDY MATERIAL

M. COM.

PAPER - 14

Advanced Statistical Concepts & Tools

POST GRADUATE COMMERCE

PREFACE

In the curricular structure introduced by this University for students of Post-Graduate degree programme, the opportunity to pursue Post-Graduate course in Subjects introduced by this University is equally available to all learners. Instead of being guided by any presumption about ability level, it would perhaps stand to reason if receptivity of a learner is judged in the course of the learning process. That would be entirely in keeping with the objectives of open educaion which does not believe in artificial differentiation.

Keeping this in view, study materials of the Post-Graduate level in different subjects are being prepared on the basis of a well laid-out syllabus. The course structure combines the best elements in the approved syllabi of Central and State Universities in respective subjects. It has been so designed as to be upgradable with the addition of new information as well as results of fresh thinking and analysis.

The accepted methodology of distance education has been followed in the preparation of these study materials. Co-operation in every form of experienced scholars is indispensable for a work of this kind. We, therefore, owe an enormous debt of gratitude to everyone whose tireless efforts went into the writing, editing and devising of proper lay-out of the meterials. Practically speaking, their role amounts to an involvement in invisible teaching. For, whoever makes use of these study materials would virtually derive the benefit of learning under their collective care without each being seen by the other.

The more a learner would seriously pursue these study materials the easier it will be for him or her to reach out to larger horizons of a subject. Care has also been taken to make the language lucid and presentation attractive so that may be rated as quality self-learning materials. If anything remains still obscure or difficult to follow, arrangements are there to come to terms with them through the counselling sessions regularly available at the network of study centres set up by the University.

Needless to add, a great part of these efforts is still experimental—in fact, pioneering in certain areas. Naturally, there is every possibility of some lapse or deficiency here and there. However, these do admit of rectification and further improvement in due course. On the whole, therefore, these study materials are expected to evoke wider appreciation the more they receive serious attention of all concerned.

Prof. (Dr.) Subha Sankar Sarkar
Vice-Chancellor

First Reprint : November, 2017

Printed in accordance with the regulations of the Distance Education Bureau of The University Grants Commission.

POST-GRADUATE : Commerce
[M. Com.]

Paper - 14

Module - 1 & 2

Advanced Statistical Concepts & Tools

Course Writing
Dr. Jadav Krishna Das

Netaji Subhas
Open University

Post Graduate
Commerce
M.Com-14

Module 1

Module 2

Unit 1 ◻ Basic Concept of Probability

Structure

1.1 Introduction

Chance is what makes life worth living. If anything was known in advance, imagine the disappointment, if decision makers had perfect information about the future as well as the present and the past there would be no need to consider the concept of probability.

However, it is unusually the case that uncertainty can not be eliminated and hence its presence should be recognised and used in the process of decision making. Information about uncertainty is often available to the decision maker in the form of probabilities. There are many events which are associated with our real life where the results can not be predicted with certainty. For example :

a) A sales manager can not say with certainty that he will achieve the sales target in a month.

b) The future life of an electric bulb or tube can not be predicted in advance. A producer can not ascertain the future demand for his product with certainty.

Such phenomena are frequently observed in Economics, business and social sciences or even in our day-to-day life.

1.2 Basic Terminology

Random Experiment : An experiment is an activity and it is called a random experiment if when conducted repeatedly under essentially homogeneous conditions, the result is not unique but may be any one of the outcomes. It is called random because it depends upon chance. For example: tossing a coin, throwing a die, drawing cards from a full pack, drawing a ball from a box etc.

Outcome: The result of a random experiment will be called an outcome.

Trial and Event : Performing a random experiment is called a trial and out comes are termed as events. For example : throwing a die is a trial and getting any one of the faces 1,2,3, ...6 is an event.

Simple and compound event: Elementary or simple events are those which can not be decomposed further. Compound events, on the other hand, can be decomposed into several simple events.

Exhaustive event: The total number of possible elementary events associated with a random experiment is known as exhaustive events. For example, if two coins are tossed the exhaustive events will be HH, HT, TH, TT.

Mutually Exclusive Events: By mutually exclusiveness of events is meant that the simultaneous occurrence of those events is not possible.

For example: In a coin tossing experiment head and tail cannot occur simultaneously. Thus head and tail are mutually exclusive events.

Equally likely events: The events are said to be equally likely or equally probable if none of them is expected to occur in preference to the other. In tossing a coin all the outcomes, viz., H, T are equally likely if the coin is an unbiased one.

Independent events: If the occurrence of one event does not affect the probability of the occurrence or non-occurrence or of other events, then events are said to be independent of each other.

Favourable cases to an event : The number of outcomes of a random experiment which result in the happening of the event concerned are known as the cases favourable to the event.

For example, in drawing a card from a full pack of cards, the number of cases? favourable to drawing a spade is 13 because there are 13 spades in a full pack of cards.

Sample space: Each concfeivable outcome of a random experiment under consideration is said to be a sample point. The totality of all conceivable sample points is called a sample space Sample 3 pace of a trial conducted by 3 tossing of a coin is {HHH, HHT. HTH, THH, TTH, THT, HTT and TTT} i.e.; 8 sample points constitute the sample space.

1.3 Classical or a-Pripri Definition of Probability

If a random experiment results in N exhaustive, mutually exclusive and equally likely outcomes and M of them are favourable to an event E, then the probability of the event E, denoted by P(E), is defined as.

$$P(E) = M/N.$$

Remarks : Since, M and N are non-negative integers with

$$0 \leq M \leq N$$

i.e. $0 \leq M/N \leq 1$

i.e. $0 \leq P(E) \leq 1$, for any event E.

If P(E) = 0, then E is called an impossible event and is denoted by ϕ

If P(E) =1, then E is called a sure event and is denoted by S.

Limitations : The classical definition of probability suffers from the following limitations :

1. This definition will not work if the size of the sample space is infinite i.e. $N \to \infty$.
2. The definition is true when the elementary events are equally likely. So the definition is circular in nature.
3. It is applicable if the outcomes are exhaustive, mutually exclusive and equally likely, without which the definition is inapplicable.
4. The definition has limited application in the games, viz., coin tossing, die throwing cases etc. Practically it has limited application in the prospective field of Statistics and probability.

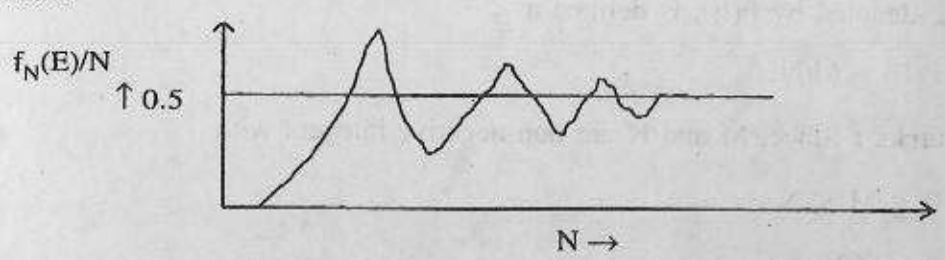
1.4 Empirical or Statistical Definition of Probability

Let $f_N(E)$ be the number of times an event E occurs in N repetition of a random experiment under essentially homogeneous condition. The ratio $f_N(E)/N$ gives the relative frequency of the event E. Then the probability of the occurrence of the event E, denoted by P(E), is given by

$$P(E) = \lim_{N \to \infty} \frac{f_N(E)}{N},$$

provided the limit exists.

It can be demonstrated graphically as follows. Along the horizontal axis we measure the number of times the coin is tossed and on the vertical axis relative frequency is measured.



As $N \to \infty$, the probability of occurrence of H or T when an unbiased coin is tossed is just 0.5.

Note : 1. In this definition we require to repeat the random experiment for an infinitely large number of times under homogeneous and identical conditions..

2. Compared to the classical definition this definition is broad enough in the sense that it is applicable if the sample space contains infinite number as well as not equally likely event points.

3. This definition gives a definite operational meaning of probability.

Limitations

1. The conditions may not remain identical, specially when the number of trial is very large.
2. The relative frequency may not attain a unique value, no matter what ever large N may be.
3. It may not be possible to repeat an experiment a large number of times.
4. Like the clasical definition, this definition does not lead to any mathematical treatment of probability.

1.5 Axiomatic or Modern Definition of Probability

Let S be a sample space of a random experiment. If to each event; E $(E \subseteq S)$ we associate a real number P(E), then P(E) is called the probability of event E, if the following axioms are satisfied :

Axiom 1. For any event $E \subseteq S$

$$P(E) \geq 0.$$

Axiom 2. For the sample sapce S, P(S) = 1.

Axiom 3. For any finite or countably infinite number of mutually exclusive events $E_1, E_2, E_3, \ldots$ of S

$$P(E_1 \cup E_2 \cup E_3 \cup \ldots) = P(E_1) + P(E_2) + P(E_3) + \ldots$$

Note: It may be noted that the axiomatic definition of probability is a general case which includes the classical and the statistical definitions as its particular cases. Besides this, it gives a number of mathematical rules that are useful for further mathematical treatment of the subject of probability.

1.6 Set Theory

Definition of a Set: .A well defined collection of distinct objects is called a set. The objects may be people, number, books, number of telephone calls etc.

Example.

1. The set of rivers in West Bengal.
2. The set of students in a Management College.
3. The set of consonants in the English Alphabet.

Let us suppose that 1,2,3,4,5,6,7 are the elements of a set A. Then we represent the set A by Roster method and we write A = { 1, 2, 3,4, 5, 6, 7}.

Element of a Set :

If 'a' is an element of a set A, we write it as $a \varepsilon A$

which is read as a belongs to A or a is in A.

Null Set :

A set which contains no element is called the *null set* or empty set. It is denoted by ϕ. If A is a null set, then A = { . }.

Singleton Set :

A set which contains only one element is called a *singleton* or *unit set.*

A = { I } is a singleton set.

Sub-set :

If every element of a set A is also an element of a set B, then A is said to be a *sub-set* of B and we write

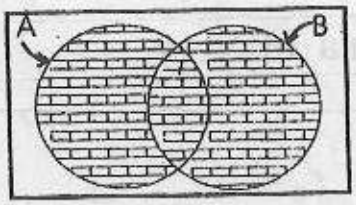
$A \subseteq B$ or $B \supseteq A$ % Here A is a sub-set of B while B is a superset of A.

Universal Set :

In any application of set theory, all the sets under investigation are likely to be considered as sub-sets of a particular set. This set is called the *Universal Set.* It is denoted by U or S.

Union of Sets :

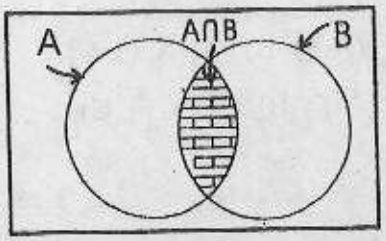
The union of sets A and B is the set of all elements which belong to either in A or in B or in both A and B. It is written as A $\cup$ B. It has been shown interms of the following venn diagram.



A $\cup$ B = B $\cup$ A = The shaded region.

Intersection of sets :

The intersection of sets of A and B is the set of all elements which belong to both A and B. It is denoted by A $\cap$ B. It has been demonstrated in the following venn diagram.

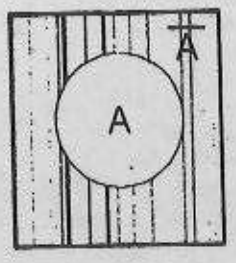


A $\cap$ B = B $\cap$ A = The shaded region.

Complement of a Set :

The complement of a set A is the set of all elements of the universal set U which do not belong to A. This is denoted by $\bar{A}$ or A^C or A'.

It has been shown in the following venn diagram.



Difference of Two Sets :

The difference of two sets A and B is the set of all elements which belong to A but do not belong to B. This is written as A – B.

1.7 Laws of Algebra of Sets

1. **Commutative Laws :**
   - i) $A \cup B = B \cup A$.
   - ii) $A \cap B = B \cap A$.
2. **Associative Laws :**
   - i) $A \cup (B \cup C) = (A \cup B) \cup C$.
   - ii) $A \cap (B \cap C) = (A \cap B) \cap C$.
3. **Distributive Laws :**
   - i) $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.
   - ii) $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.
4. **De Morgan's Laws :**
   - i) $(A \cup B)' = A' \cap B'$.
   - ii) $(A \cap B)' = A' \cup B'$.
5. **Idempotent Laws :**
   - i) $A \cup A = A$.
   - ii) $A \cap A = A$.
6. **Identity Laws :**
   - i) $A \cup \phi = A$
   - ii) $A \cap \phi = \phi$
   - iii) $A \cup S = S$
   - iv) $A \cap S = A$

7. Complement Laws :

i) $A \cup A' = S$

ii) $A \cap A' = \phi$

iii) $(A')' = A$

iv) $S' = \phi$ and $\phi' = S$.

The various operations on sets, viz., union, intersection, difference and complementation can be explained through examples below :

Example :

Let $S = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$,

$A = \{1, 2, 3, 4, 5, 6\}$ and

$B = \{5, 6, 7, 8, 9\}$.

Then $A \cup B = B \cup A = \{1, 2, 3, 4, 5, 6, 7, 8, 9\} = S$,

$A \cap B = B \cap A = \{5, 6\}$,

$\bar{A}$ or A^C or $A' = \{7, 8, 9\}$,

$A \cap B' = A - B = \{1, 2, 3, 4\}$ and

$B \cap A' = B - A = \{7, 8, 9\}$.

1.8 Some Fundamental Theorems on Probability

Theorem 1 :

$P(\phi) = 0$, when ϕ is an impossible event.

Proof :

From the algebra of sets

$A \cap \phi = \phi$.

That is, A and ϕ are mutually exclusive.

Now $A \cup \phi = A$.

So, by Axiom 3

$P(A) = P(A \cup \phi) = P(A) + P(\phi)$.

Hence, $P(\phi) = 0$.

Theorem 2 : For any event A,

$P(A^C) = 1 - P(A)$.

Proof : $A \cap A^C = \phi$,

that is, A and A^C are mutually exclusive events.

Further $A \cup A^C = S$.

Hence by Axiom 2 and 3,

$1 = P(S) = P(A \cup A^c) = P(A) + P(A^C)$.

Therefore, $P(A^C) = 1 - P(A)$.

Theorem 3 : If $A \subseteq B$ for any two events A and B in S,

Then $P(A) \leq P(B)$.

Proof : If $A \subseteq B$, then B can be written as

$B = A \cup (A^C \cap B)$

Where A and $(A^C \cap B)$ are mutually exclusive.

Hence by Axiom 3,

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

Further, $P(A^C \cap B) \geq 0$, by Axiom 1.

Hence $P(B) \geq P(A)$.

Theorem 4 : $0 \leq P(A) \leq 1$.

Proof : Since $A \subseteq S$, using theorem 3

$P(A) \leq P(S)$.

i.e. $P(A) \leq 1$.

Again, by Axiom 1, $P(A) \geq 0$.

Hence, $0 \leq P(A) \leq 1$.

Theorem 5 : For any two events A and B in S

(i) $P(A) = P(A \cap B) + P(A \cap B^C)$

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

Proff : Events $A \cap B$ and $A \cap B^C$ are mutually exclusive. Again,

$A = (A \cap B) \cup (A \cap B^C)$.

Hence by Axiom 3,

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

Similarly, (ii) can be proved.

Theorem 6 : For any two events A and B in S

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

Proof :

One can write $A \cup B$ as

$A \cup B = (A \cap B^C) \cup (A \cap B) \cup (A^C \cap B)$

where the events $(A \cap B^C)$, $(A \cap B)$ and $(A^C \cap B)$ are mutually exclusive.

Hence, by Axiom 3,

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

Now from Theorem 5

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

and $P(A^C \cap B) = P(B) - P(A \cap B)$

Substituting these values,

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

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

Corollary : If A and B are mutually exclusive then

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

Theorem 7 : For any three events A, B and C in S,

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

Proof :

In Theorem 5 if B is replaced by $(B \cup C)$ then

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

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

$= P(A) + P(B \cup C) - P[(A \cap B) \cup (A \cap C)]$, by the distributive law.

So repeated application of Theorem 6 gives.

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

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

Hence the proof of the theorem.

Theorem 8 :

For any n events $B_1, B_2, B_3, \ldots\ldots, B_n$ in S

$$P\left(\bigcup_{i=1}^{n} B_i\right) = \sum_{i=1}^{n} P(B_i) - \sum_{i,\, j=1,\, i<j}^{n} P(B_i \cap B_j) + \sum_{i,\, j,\, k=1,\, i<j<k}^{n} P(B_i \cap B_j \cap B_k) - \ldots\ldots + (-1)^{n-1} P\left(\bigcap_{i=1}^{n} B_i\right).$$

Corollary : If $B_1, B_2, \ldots\ldots, B_n$ are mutually exclusive events, then

$$P\left(\bigcap_{i=1}^{n} B_i\right) = \sum_{i=1}^{n} P(B_i).$$

1.9 Some Important Results

1. For any two events A and B in S,

   $P(A \cap B) \leq P(A)$ or $P(B) \leq P(A \cup B) \leq P(A) + P(B)$.

2. For any n events $B_1, B_2, \ldots, B_n$ in S

$$P(B_1 \cup B_2 \cup \ldots \ldots \cup B_n) \leq P(B_1) + P(B_2) + \ldots . + P(B_n).$$

This result is known as Boole's Inequality.

3. $$P\left(\bigcap_{i=1}^{n} B_i\right) \geq 1 - \sum_{i=1}^{n} P(B_i^c)$$

$$P\left(\bigcap_{i=1}^{n} B_i\right) \geq \sum_{i=1}^{n} P(B_i) - (n-1).$$

This result is known as Bonferroni's Inequality

for n = 2, $P(B_1) \cap B_2) \geq P(B_1) + P(B_2) - 1$.

1.10 Conditional Probability

Let us consider two events A and B in the sample space S of a random experiment such that P(B) > 0. Then the conditional probability of A given that B has actually occurred, denoted by P(A/B), is defined by

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

Note : The conditional probability P(A/B) will be undefined if P(B) = 0.

1.11 Compound Probability

From the definition of conditional probability it follows that

$$P(A \cap B) = P(B)\, P(A / B) \text{ with } P(B) > 0.$$

The probability of occurrence of the event A as well as B is given by the product of unconditional probability of B and conditional probability of A given that B has already occurred. This is known as multiplicative *thorem* of probability.

Similarly, if $P(A) > 0$, the conditional probability of B given that A has already occurred is

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

This implies that

$$P(A \cap B) = P(A)\, P(B/A) \quad \text{with } P(A) > 0.$$

Similarly, for three events A, B and C in S

$$P(A \cap B \cap C) = P(A)\, P(B/A)\, P(C/A \cap B) \quad \text{with } P(A \cap B) = 0.$$

In general, for n events $B_1, B_2, \ldots\ldots, B_n$ in S

$$P(B_1 \cap B_2 \cap \ldots \cap B_n) = P(B_1)\, P(B_2/B_1)\, P(B_3/B_1 \cap B_2) \ldots P(B_n / B_1 \cap B_2 \ldots \cap B_{n-1})$$

with $P(B_1 \cap B_2 \ldots\ldots \cap B_{n-1}) > 0$.

1.12 Independent Events

Two events A and B in a sample space S of a random experiment are said to be independent (or statistically independent) when the occurrence of the event B does not affect the probability of occurrence of A. By notation two events A and B are said to be statistically independent if

$$P(A \cap B) = P(A)\, P(B).$$

By the compound probability theorem we have

$$P(A \cap B) = P(B)\, P(A/B), \text{ if } P(B) > 0$$

$$= P(A)\, P(B/A), \text{ if } P(A) > 0.$$

If $P(A/B) = P(A)$ and $P(B/A) = P(B)$, then

$$P(A \cap B) = P(A)\, P(B).$$

Three events A, B and C in S are independent if

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

In general, for n (finite) events $B_1, B_2, \ldots\ldots B_n$ are independent if

$$P(B_1 \cap B_2 \cap \ldots\ldots \cap B_n) = P(B_1)\, P(B_2) \ldots\ldots\ldots P(B_n).$$

Result : If A and B are two independent events, then

(i) A and B^C (ii) A^C and B (iii) A^C and B^C are also independent events.

Proof : (i) From Theorem 5

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

Or, $$P(A \cap B^C) = P(A) - P(A \cap B)$$

$$= P(A) - P(A)\,P(B) \quad \text{(since A and B are independent)}$$

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

$$= P(A)\,P(B^C).$$

Thus the events A and B^C are independent.

(ii) Similar is the proof as in (i).

(iii) For any event E

$$P(E^C) = 1 - P(E).$$

Now $$P(A^C \cap B^C) = P(A \cup B)^C \quad \text{(by Demorgan's law)}$$

$$= 1 - P(A \cup B)$$

$$= 1 - P(A) - P(B) + P(A \cap B) \quad \text{(by Theorem 6)}$$

$$= 1 - P(A) - P(B) + P(A)\,P(B) \quad \text{(Since A and B are independent)}$$

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

$$= P(A^C)\,P(B^C)$$

Therefore, the events A^C and B^C are also independent.

Note : If A and B are independent, then

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

(ii) $$P(A \cup B) = 1 - P(A \cup B)^C$$

$$= 1 - P(A^C \cap B^C)$$

$$= 1 - P(A^C)\,P(B^C)$$

1.13 Pair-Wise Independent Events

A set of n events $B_1, B_2, \ldots\ldots, B_n$ are said to be pair-wise independent if any pair of these events are independent.

That is, if $P(B_i \cap B_j) = P(B_i)\, P(B_j)$ for $i \neq j$ $(i, j = 1, 2, \ldots, n)$.

In this case we have ${}^nC_2$ conditions.

1.14 Mutually Independent Events

A set of n evens $B_1, B_2, \ldots\ldots, B_n$ are said to be mutually indepnedent if the probability of simulatancous occurrence of any number of events is equal to the product of their individual probabilities. That is, if

$P(B_i \cap B_j) = P(B_i)\, P(B_j)$ for $i \neq j$, $i, j = 1, 2, \ldots, n$

$P(B_i \cap B_j \cap B_k) = P(B_i)\, P(B_j)\, P(B_k)$ for $i \neq j \neq k$, $i, j, k = 1, 2, \ldots, n$

.....

.....

and $P(B_1 \cap B_2 \cap \ldots \cap B_n) = P(B_1)\, P(B_2) \ldots\ldots P(B_n)$

The total no, of conditions when the events $B_1, B_2, \ldots B_n$ are mutually independent are ${}^nC_2 + {}^nC_3 + \ldots + {}^nC_n = ({}^nC_0 + {}^nC_1 + {}^nC_2 \ldots + {}^nC_n) - {}^nC_1 - {}^nC_0$

$$= 2^n - {}^nC_1 - {}^nC_0 = 2^n - n - 1.$$

Obviously, mutual independence implies pair-wise independence but the converse is not necessarily true.

1.15 Bayes' Theorm (Thomas Bayes – 1763)

An event E can occur only if one of the mutually exclusive and exhaustive events $B_1, B_2, \ldots B_n$ occurs. Assume that the unconditional probabilities $P(B_1), P(B_2), \ldots$

$P(B_n)$ and the conditional probabilities $P(E/B_1)$, $P(E/B_2)$, ... $P(E/B_n)$ are known. Then the conditional probability of a specified event B_i when the event E has actually occurred is given by

$$P(B_r/E) = [P(B_r)\ P(E/B_r)] \ / \ [\ \sum_{i=1}^{n} P(B_i)P(E/B_i)], \ r = 1, 2, ... n.$$

This is known as Bayes' Theorem.

Proof : Given that $B_1, B_2, ..., B_n$ are mutually exclusive and exhaustive.

So, $B_1 \cup B_2 \cup ..... \cup B_n = S$ (sample space)

Now for any arbitrary event E we have

$$E = E \cap S$$
$$= E \cap (B_1 \cup B_2 \cup .... \cup B_n)$$
$$= (E \cap B_1) \cup (E \cap B_2) \cup ..... \cup (E \cap B_n).$$

The events $(E \cap B_1)$, $(E \cap B_2)$, ..., $(E \cap B_n)$ are also mutually exclusive.

Therefore,

$$P(E) = P(E \cap B_1) + P(E \cap B_2) + ... + P(E \cap B_n)$$
$$= P(B_1)\ P(E/B_1) + P(B_2)\ P(E/B_2) + ... + P(B_n)\ P(E/B_n)$$
$$= \sum_{i=1}^{n} P(B_i)\ P(E/B_i).$$

Again, $P(B_r \cap E) = P(B_r)\ P(E/B_r)$ if $P(B_r) > 0$.

and $P(B_r \cap E) = P(E)\ P(B_r/E)$ if $P(E) > 0$.

Hence, $P(B_r/E) = [P(B_r)\ P(E/B_r)] \ / \ P(E)$

$$= [P(B_r)\ P(E/B_r)] \ / \left[\sum_{i=1}^{n} P(B_i)\ P(E/B_i)\right]$$

Hence the theorem is proved.

1.16 Problems on Probability

Example 1 : An unbiased die is thrown and the number of points appearing on the uppermost face is noted. What is the probability of

(a) and odd number, (b) an even number and (c) more than 4.

Solution : The sample space of the random experiment shows 6 possible sample points, Viz. 1, 2, 3, 4, 5 and 6. The outcomes are mutually exclusive, exhaustive and equally likely. Total number of outcomes (N) = 6.

(a) Let E_1 denote the event that an odd number of points is obtained .Among 6 outcomes 3 (Viz. 1, 3, 5) are favourable to E_1.

Hence, P [an odd number is obtained] = $P(E_1) = 3/6 = 1/2$.

(b) Let E_2 denote the event that an even number of points is obtained. So 3 outcomes (Viz. 2, 4, 6) are favorable to E_2.

Now, P [an even number is obtained] = $P(E_2) = 3/6 = 1/2$.

(c) Let E_3 be the event that a number is obtained which is more than 4. In this case 2 outcomes (Viz. 5, 6) are favorable to E_3.

So, P (number is more than 4) = $P(E_3) = 2/6 = 1/3$.

Example 2 : An unbiased coin is tossed thrice. What fs the probability that there are a) three heads, b) at least one head and c) at most one tail?

Solution : Here the sample space S gives $2^3 = 8$ sample points which are mutually exclusive and exhaustive. S = {HHH, HTH, THH, TTH, HHT, HTT, THT, TTT}.

Since the coin is unbiased, these sample points are equally likely.

a) Let E_1 be the event of getting the three heads. There is only one sample point HHH which is favourable to the event E_1.

So, $P(E_1)=1/8$.

(b) Let E_2 be the event of getting at least one head. Out of 8 sample points 7 are favourable to the event E_2. Hence the required probability is

$P(E_2) = 7/8$.

(c) Let E_3 be the event of getting at most one tail. In this case out of 8 cases,

4 cases (Viz., HHH, THH, HTH, HHT) are favourable to the event E_3 . Hence the required probability is

$P(E_3) = 4/8 = 1/2$.

Example 3 : stockist has 30 items in a lot, out of which 8 are defective. A customer selects three items from the lot.

a) What is the probability that all 3 are defective?

b) What is the probability that out of 3 items two are non-defective and one is defective?

Solution : 3 items can be selected out of 30 items in $^{30}C_3$ ways. These cases are mutually exclusive, exhaustive and equally likely. Out of 30 items there are 8 defective items and 22 are non-defective items.

a) Let E_1 be the event of selecting all 3 non-defective items. Hence number of cases favourable to the event E_1 is $^{22}C_3$,

So, $P(E_1) = {}^{22}C_3 / {}^{30}C_3 = (22 \times 21 \times 20) / (30 \times 29 \times 28) = 0.3793$

b) Let E_2 be the event of getting 2 non-defective items and one defective item. So the customer has to select 2 non-defective items from 22 items and 1 defective item from 8 defective items. So the no. of cases favourable to the event E_2 is $^{22}C_2 \times {}^{8}C_1$. Hence the required probability is $P(E_2) = {}^{22}C_2 \times {}^{8}C_1$, $/ {}^{30}C_3 = (3 \times 22 \times 21 \times 8) / (30 \times 29 \times 28) = 0.4552$.

Example 4 : The probability that a contractor gets a plumbing contract is 2/3 and the probability that he will not get an electric contract is 5/9. If the probability of getting at least one contract in 4/5, what is the probability that he will get both the contracts?

Solution : Let E_1 be the event that the contractor will get the plumbing contract and E_2 be the event that the contractor will get the electric contract.

The probability of getting least one contract is = $P(E_1 \cup E_2) = 4/5$

The probability of getting the plumbing contract is = $P(E_1) = 2/3$

The probability of getting the electric contract is = $P(E_2) = 1 - P(E_2^C)$

$= 1 - 5/9 = 4/9$

Hence the probability of getting both the contracts is $P(E_1 \cap E_2)$. This can be obtained by the theorem of total probability as follows:

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

Or, $P(E_1 \cap E_2) = P(E_1) + P(E_2) - P(E_1 \cup E_2)$.

$= 2/3 + 4/9 - 4/5 = 14/45$.

Example 5 : In a bridge game, North and South have 9 spades between them. Find the probability that either East or West has no Spades.

Solution : The players are designated by the position they have occupied , Viz. North, South, East and West.

Let A denote the event that East has no spades and B denote the event that West no spades.

There are only 13 spades in a pack of 52 cards and each player has 13 cards.

Since South and North have 9 spades between them, East and West have 4 spades between them.

It is clear that East and West have 26 cards between them in total, of which 4 are spades.

Hence, P [either East or West has no spades.] = $P[A \cup B]$

$= P(A) + P(B) - 2\ P(A \cap B)$.

Here the events A and B are mutually exclusive. So, $P(A \cap B) = 0$.

i.e., $P(A \cup B) = P(A) + P(B) = 2P(A) = 2P(B)$.

Now, $P(A) = P(B) = {}^{22}C_{13} / {}^{26}C_3 = 11/230$.

Therefore, $P(A \cup B) = 2 \times (11/230) = 11/115$.

Example 6 : A candidate is selected for interview for 3 posts. For the first post there are 3 candidates. For the second post there are 4 candidates. For the third post there are 2 candidates. What is the chance of getting at least 1 post?

Solution : Let E_1, E_2 and B_3 denote the respective events that the candidate will get the first, second and the third post. In the question it is given that

$P(E_1) = 1/3$, $P(E_2) = 1/4$ and $P(E_3) = \frac{1}{2}$.

Here the events E_1 E_2 and E_3 are mutually independent.

Hence, P [The chance of getting at least one post]

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

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

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

(Since the events are independent)

$= 1/3 - 1/4 - 1/2 - (1/3) \times (1/4) - (1/3) \times (1/2) - (1/4)(1/2) + (1/3) \times (1/4) \times (1/2)$

$= 3/4.$

Alternatively, $P(E_1 \cup E_2 \cup E_3) = 1 - P(E_1 \cup E_2 \cup E_3)^c$

$= 1 - P(E_1^C \cap E_2^C \cap E_3^C)$

$= 1 - P(E_1^C)\, P(E_2^C)\, P(E_3^C)$

$= 1 - (1-1/3)\,(1-1/4)\,(1-1/2) = 3/4.$

Example 7: A bag contains 8 red and 5 white balls. Two successive draws of 3 balls are made without replacement. Find the probability that the first drawing will give 3 white balls and the second 3 red balls.

Solution : Let E_1 be the event that the first drawing gives 3 white balls and E_2 be the event that the second drawing gives 3 red balls. We are required to calculate $P(E_1 \cap E_2]$ $E_2)$. By the compound law of probability

$$P(E_1 \cap E_2) = P(E_1)\, P(E_2/E_1).$$

To find $P(E_1)$, the total number of possible outcomes is $13_{C_3}|$ and it is clear that only ${}^5C_3$ cases are favourable to the event E_1.

Hence, $P(E_1) = {}^5C_3 / {}^{13}C_3 = 5/143.$

Again, to compute $P(E_2/E_1)$, it is to be considered that 3 white balls are already selected. So there are 8 red balls and 2 white balls in the bag. Hence the probability of getting 3 red balls in the second draw is

$$P(E_2 /E_1) = {}^8C_3 / {}^{10}C_3 = 7/15.$$

Therefore , the required probability is

$P(E_1 \cap E_2) = P(E_1)P(E_2/E_1) = (5/143) \times (7/15) = 7/429.$

Example 8 : Three salesmen A, B and C have been given a target of selling 10,000 units of a particular product, the probability of their success being respectively 0.25, 0.30 and 0.40. If those three salesmen try to sell the product , find the probability

that there is success of only one and failure of the other two.

Solution: Let E_1, E_2 and E_3 denote respectively the events that A, B and C will be successful to achieve a target of selling 10,000 units.

Here it is given that

$P(E_1) = 0.25$, $P(E_2) = 0.30$ and $P(E_3) = 0.40$.

Hence, $P(E^C_1) = 0.75$, $P(E^C_2) = 0.70$ and $P(E^C_3) = 0.60$.

If E denotes the event that success of one salesman and failure of the other two, then E can occur in the following mutually exclusive ways:

Success of A and failure of B and C.

Success of B and failure of A and C.

Success of C and failure of A and B.

i.e. $E_1 \cap E^C_2 \cap E^C_3$ or $E^C_1 \cap E_2 \cap E^C_3$ or $E^C_1 \cap E^C_2 \cap E_3$.

Hence,

$P(E) = P[(E_1 \cap E^C_2 \cap E^C_3) \cup E^C_1 \cap E_2 \cap E^C_3) \cup (E^C_1 \cap E^C_2 \cap E_3)]$

That is, $P(E) = P(E_1 \cap E^C_2 \cap E^C_3) + P(E^C_1 \cap E_2 \cap E^C_3) + P(E^C_1 \cap E^C_2 \cap E_3)$

$= P(E_1) P(E^C_2) P(E^C_3) + P(E^C_1)P(E_2)P(E^C_3) + P(E^C_1)P(E^C_2)P(E_3)$

(Since E_1, E_2 and E_3 are mutually independent)

$= 0.25 \times 0.70 \times 0.60 + 0.75 \times 0.30 \times 0.60 + 0.75 \times 0.70 \times 0.40.$

$= 0.45$

Example 9 : An electronic device is made up of three components A, B and C. The probability of failure of the component A is 0.01, that of B is 0.05 and of C is 0.2 in some fixed time period. Find the probability that the device will work satisfactorily during the period of work. Assume that the three components work independently of one another.

Solution : Let E_1, E_2 and E_3 denote respectively the events that the failure of components A, B and C are in some fixed period of time. Given that

$P(E_1) = 0.01$, $P(E_2) = 0.05$ and $P(E_3) = 0.02$.

Again, the components are working independently.

So the probability that the device will work satisfactorily during the given period of time will be

Probability that all the components are working simultaneously.

$= P(E^C_1 \cap E^C_2 \cap E^C_3)$.

$= P(E^C_1)P(E^C_2)P(E^C_3)$.

$= (1-0.01)\ (1-0.05)\ (1-0.02)$

$= 0.99 \times 0.95 \times 0.98$.

$= 0.92169$.

Example 10: Suppose that there is a chance for a newly constructed house to collapse whether the design is faulty or not. The chance that the design is faulty is 0.1.The chance that the house collapses if the design is faulty is 0.95 and otherwise it is 0.45. It is seen that the house collapsed. What is the probability that it is due to faulty design?

Solution : Let B_1 and B_2 denote the events that the design is faulty and the design is not faulty. Let A be the event denoting that the house is collapsed. We are given that

$P(B_1) = 0.1$ and $P(B_2) = 1-P(B_1) = 0.9$

$P(A/B_1) = 0.95$ and $P(A/B_2) = 0.15$.

Here we are interested in computing the probability that the design is faulty given that the house collapsed. By using Bayes' theorem we get

$P(B_1/A) = [P(B_1)\ P(A/B_1)]\ /\ [P(B_1)\ P(A/B_1) + P(B_2)\ P(A/B_2)]$

$= 0.1 \times 0.95\ /\ [0.1 \times 0.95 + 0.9 \times 0.45]$

$= 0.19$

Example 11 : A company has 3 plants to manufacture 10,000 scooters in a month. Plant I, II and III manufacture 5000, 3000 and 2000 respectively per month. 90% 92% and 95% scooters are rated standard quality in Plant I, II and III respectively.

i) What is the probability that a scooter selected at random is of standard quality?

ii) What is the probability that a scooter selected at random comes from plant II if it is known that the scooter is of standard quality?

Solution : Let us define the following events :

E_1 : A scooter is manufactured by plant I.

E_2 : A scooter is manufactured by plant II,

E_3 : A scooter is manufactured by plant III.

E : A scooter is rated as standard quality.

It is given that

$P(E_1) = 5000/10000 = 0.5$, $P(E_2) = 3000/10000 = 0.3$,

$P(E_3) = 2000/10000 = 0.2$, $P(E/E_1) = 0.90$,

$P(E/E_2) = 0.92$ and $P(E/E_3) = 0.95$,

i) The probability that a scooter selected at random is of standard quality = P(E)

$= P(E_1)\, P(E/E_1) + P(E_2)\, P(E/E_2) + P(E_3)\, P(E/E_3)$.

$= 0.5 \times 0.9 + 0.3 \times 0.92 + 0.2 \times 0.95$

$= 0.8455$

ii) The probability that a scooter selected at random comes from Plant II if it is known that the scooter is of standard quality = $P(E_2/E)$

By Bayes' theorem,

$P(E_2 / E) = P(E_2)\, P(E/E_2) / P(E)$

$= 0.3 \times 0.92/0.8455$

$= 0.3264$

Example 12: Two players A and B alternatively toss an unbiased coin. He who first tosses a head wins the game. If A starts, find their respective probability of winning the game.

Solution : Let E_1 and E_2 denote the events that A and B toss a head respectively. Obviously, the two events are independent.

Since A starts , he can first obtain a head in the following mutually exclusive ways :

(i) E_1 occurs only, (ii) $E^C_1 \cap E^C_2 \cap E_1$ occurs, (iii) $E^C_1 \cap E^C_2 \cap E^C_1 \cap E^C_2 \cap E_1$ occurs and so on.

Hence, $P(E^C_1) = P(E^C_2) = 1/2$.

Therefore, the probability that A wins the game

$$= P(E_1) + P(E^C_1 \cap E^C_2 \cap E_1) + P(E^C_1 \cap E^C_2 \cap E^C_1 \cap E^C_2 \cap E_1) + \ldots$$

$$= P(E_1) + P(E^C_1)P(E^C_1)P(E_1) + P(E^C_1)P(E^C_2)P(E^C_1)P(E^C_2)P(E_1) + \ldots\ldots$$

$$= 1/2 + 1/2 \times 1/2 \times 1/2 + 1/2 \times 1/2 \times 1/2 \times 1/2 \times 1/2 + \ldots$$

$$= 1/2 + (1/2)^3 + (1/2)^5 + \ldots$$

$$= (1/2) / [1 - (1/2)^2] = 2/3.$$

So the probability that B wins the game is (1-2/3) = 1/3.

1.17 Exercises

1. What do you understand by the term probability? Discuss its importance in business decision making.
2. Explain various approaches to probability. Are they contradictory?
3. Define independent and mutually exclusive events. Can two events be mutually exclusive and independent simultaneously? Give an example to illustrate your answer .
4. Explain the meaning of conditional probability of an event. State the multiplicative rule of probability.
5. State and prove Bayes' theorem on conditional probability.
6. What are the limitations of the classical definition of probability?
7. Distinguish between pair-wise independent events and mutually independent events.
8. For any two events A and B prove that

   $P(A \cap B) \leq P(A) \leq P(A \cup B) \leq P(A) + P(B)$.
9. If A and B are two independent events, prove that i) A^C and B^C, ii) A^c and B and iii) A and B^C are also independent.

10. The eight possible outcomes e_i, i = 1,2, 8 of a random experiment are equally likely. Suppose the events A, B and C are defined as follows :

$A = \{ e_1, e_2, e_3, e_4\}$, $B = \{e_3, e_4, e_5, e_6\}$, $C = \{e_3, e_4, e_7, e_8\}$. Examine the type of independence of events A, B and C.

11. 3 compressors used in refrigerators are manufactured by a company at 3 factories located at X, Y and Z. It is known that the factory A produces twice as many compressors as factory B which produces the same number as the factory C. It is known that 2%, 3% and 4% compressors produced by factory A, B and C respectively are defective. A quality control engineer while maintaining a refrigerator finds a defective compressor. What is the probability that factory A is not to be blamed?

12. The probabilities of X, Y and Z becoming managers are 4/9, 2/9 and 1/3 respectively. The probability that a new product will be introduced if X, Y and Z become managers are 3/10, 1/4 and 4/5 respectively.

    a) What is the probability that a new product will be introduced?

    b) What is the probability that the manager appointed was Z given that the new product had been introduced?

13. The probability that a man will be alive for one year is 3/5 and the same for his wife is 4/5. Find the probability that after one year

    i) both will be alive.

    ii) only the man will be alive.

    iii) only the wife will be alive.

    iv) at least one of them will be alive.

14. Four cards are drawn one by one from a full pack of 52 cards. What is the probability that they belong to (i) 4 different suits and (ii) different suits and denominations.

15. In a city 3 daily newspapers X, Y and 2 are published. 40% of the people of the city read X, 50% read Y, 30% read Z, 20 % read both X and Y, 15% read X and Z, 10% read Y and Z and 24 % read all the 3 papers. Calculate the percentage of people who do not read any one of the 3 news papers.

16. There are six hotels in a town. If 4 men check into hotels a day, what is the probability that each checks into a different hotels?

17. If P(A) = 1/4, P (B) = 2/5, P (A $\cap$ B) = 1/7, find (i) P (A $\cup$ B), (ii) P (A' $\cap$ B'), (iii) P (A $\cap$ B') and (iv) P (A' $\cup$ B').

18. If A and B are two events, prove that (i) P(A/B) < {1 - P(A')/P(B) and (ii) P(A/B) = 1 - P(A'/B), given P(B) > 0.

19. A speaks truth in 80% cases and his friend B speaks lie in 30% cases. In what percentage of cases are they likely to contradict each other in narrating the same incident?

20. A person fails to remember the last digit of the telephone number of his friend. What is the probability that at most 3 attempts are necessary to find the actual number?

21. A company has a security system comprising four electronic devices (A, B, C and D) which operate independently. Each device has a probability of 0.2 of failure. The four electronic devices are arranged so that the whole system operates if at least one of A or B functions and at least one of C or D functions. What is the probability that the whole system will fail?

22. There are 10 electric bulbs in the stock of a shop, out of which 4 are defective. A customer demands 2 bulbs and the shopkeeper picks up two bulbs randomly. What is the probability that both these bulbs are defective.?

23. The probability that a salesman of vacuum cleaner will succeed in persuading a customer on the first call is 0.6. If he fails, the probability of success on the second call is 0.3 . If he fails on the first two calls, the probability of success on the third and the last call is 0.1. Find the probability that the salesman makes a sale of vacuum cleaner to a customer.

24. Two dice are thrown at a time and outcomes are noted down . Define the events A, B and C as

    A = odd number on the first die

    B = even number on the second die.

    C = combined odd sum.

Verify that A, B and C are pair-wise independent, but they are not mutually independent.

25. A company has four production sections A, B, C and D which contribute 30%, 20%, 28% and 22% respectively to the total output. It was observed that these sections produced 1%, 2%, 4% and 6% defective items respectively. If an item is selected at random and found to be defective, what is the probability that it has come from D?

26. A and B stand in a line at random with 10 other persons. What is the probability that there are 3 persons between A and B?

27. A and B stand in a line with 10 other persons. Find the probability that there are 3 persons between A and B.

28. A fair coin is tossed six times. What is the probability of (i) exactly two heads and (ii) at least two heads?

29. Three candidates Mr. Wiseman, Miss Drinkwater and Mr. Page stand for student President in a university. A public opinion poll shows their chances of winning as 0.5, 0.3 and 0.2 respectivly. The probabilites that they will promote "student power" if they are elected are 0.7, 0.6 and 0.9 respectively. What is the probability that the "student power" will be promoted after the election?

30. Suppose A and B are two events. Show that the probability that exactly one of the events occurs is equal to $P(A) + P(B) - 2P(A \cap B)$.

Unit 2 ❑ Random Variable and its Probability Distribution

Structure

2.1 Introduction

In Probability theory the basic entities are elementary events (sample points) in the sample space. These types of entities are abstract in nature. It is not convenient to develop mathematical theory of these abstract entities. But if we have a set of real numbers we can easily develop mathematical theory. So if it is possible to associate a real number with each sample points in the sample space a mathematical theory of probability can be easily developed. For each sample point in a sample space of a random experiment we assign a real number and such assignment gives a function on the sample space which is called a random variable. This chapter deals with the concept of a random variable, its probability distribution, mathematical expectation, variance, different types of moments and moment generating function. This chapter also contains the bivariate probability distribution of two dependent random variables.

2.2 Random Variable

In tossing a coin we may associate the number 1 with the appearance of a head and the number 0 with the appearance of a tail. In throwing a die, we may associate six numbers 1, 2, 3, 4, 5 and 6 corresponding to the face that appears uppermost.

Definition 1. A real-valued function defined on the sample space is called a random variable i.e., different values of a random variable are obtained by associating a real number with each element at the sample space. A random variable is also called a stochastic variable or a chance variable or a probability variable.

Example 1. 3 coins are tossed simultaneously with the sample space of the random experiment. What are the possible values of the random variable 'number of heads' obtained?

Solution : In the random experiment of tossing 3 coins, the sample space will be

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

If X be the random variable denoting the 'number of heads', then we can assign a number to each sample point as follows :

X (HHH) = 3, X (HHT) = 2, ... X (TTT).= 0. So the range of X, R = {0, 1, 2, 3}.

In the above example the outcome set or range is discrete and then the variables are defined on a discrete set. Such variables are called discrete random variables. Some more examples are number of accidents, number of printing mistakes, number of defectives, number of telephone calls etc. But, for example, consider the life time of an electrical item, length of cloth manufactured in a textile mill, the time between arrivals of customers at a service station, consumption of petrol etc. which are defined on a continuous set and hence the variable associated with those is a continuous random variable.

If the range of a random variable has only a finitely many values or countable many values, then the random variable is called a discrete one. On the other hand, if the range of a random variable has uncountable many values over an interval, the random variable is called a continuous one.

2.3 Probability Distribution

A probability distribution associated with a discrete random variable is referred to as a probability mass function (p.m.f.) and a probability distribution associated with a continuous random variable is referred to as a probability density function (p.d.f.).

If X is a discrete random variable then the function given by $p(x) = P[X = x]$ for each x within the range of X is called the probability distribution of X.

Definition 2. The function p(x) is called the probability mass function (p.m.f.) of the discrete random variable X if and only if its values satisfy the conditions :

1. $p(x) \geq 0$, for each value within its domain.

2. $\sum_x p(x) = 1$ where the summation extends over all the values within its domain.

Example 2. Find the probability distribution of the random variable in Example 1 with the assumption that the coin is unbiased.

Soiution : Let X bc the number of heads obtained by tossing 3 unbiased coins simultaneously.

Here the range of X, $R = \{0, 1, 2, 3\}$

$P(X = 0) = p(0) = 1/8$

$p(X = 1) = p(1) = 3/8$

$P(X = 2) = p(2) = 3/8$

$P(X = 3) = p(3) = 1/8.$

The set of all possible values of the random variable X along with their respective probability is known as the probability distribution of X. So in this case the probability distribution of X can be written as :

x	:	0	1	2	3	Total
P (x)	:	1/8	3/8	3/8	1/8	1

If X is a continuous random variable, then a function with values f(x), defined over the set of all real numbers, is called the probability density function of x if

$$P[a \leq x \leq b] = \int_a^b f(x)dx,$$

for any real constants a and b with $a \leq b$.

Definition 3. The function f(x) is called the probability density function (p.d.f.) of a continuous random variable X if its values satisfy the conditions :

1. $f(x) \geq 0$ for $-\infty < x < \infty$

2. $\int_a^b f(x)dx = 1$

Example 3. Can the following be a probability density function?

$f(x) = 1/4$, if $-2 < x < 2$

$= 0$, otherwise.

If so, evaluate a) $P[x < 1]$, b) $P[|x| < 1]$ and c) $P[2x + 4 > 5]$

Solution : Here $f(x) \geq 0$, for all x.

Again, $\int_{-\infty}^{\infty} f(x)dx = \int_{-2}^{2} \frac{1}{4} dx = \frac{1}{4}[x]_{-2}^{2} = \frac{2-(-2)}{4} = \frac{4}{4} = 1.$

So f(x) satisfies both the conditions of being the pdf.

Hence, f(x) is a probability density function .

Now,

a) $P(x < 1) = \int_{-\infty}^{1} f(x)dx = \int_{-2}^{1} \frac{1}{4} dx = \frac{3}{4}$

b) $P[|X|<1] = P[-1 < x < 1]$

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

c) $P[2x + 4 > 5]$

$= P[2x > 1] = P[x > 0.5]$

$= \int_{0.5}^{\infty} f(x)dx = \frac{1}{4}\int_{0.5}^{2} dx = \frac{1.5}{4} = \frac{3}{8}.$

2.4 Cumulative Distribution Function

There are many situations in which we are interested in knowing the probability that the value of a random variable is less than or equal to some real number x. So, let us write the probability that X takes a value less than or equal to x as $F(x) = P[X \leq x]$ and refer to this function defined for all real numbers of x as the distribution function or the cumulative distribution function of X.

Definition 4. If X is a discrete random variable, the function given by

$$F(x) = P[X \leq x] = \sum_{u \leq x} p(u)$$

where p(u) is the value of the p.m.f. of X at u. F(x) is called the distribution function or cumulative distribution function (c.d.f.) of the random variable X.

Distribution function F(x) satisfies the following properties:

1) $0 \leq F(x) \leq 1$
2) $F(-\infty) = 0$
3) $F(\infty) = 1$
4) $F(a) \leq F(b)$ for all $a \leq b$.
5) F(x) is right continuous.

If X is a continuous random variable, then the cumulative distribution function is given by $F(x) = P[X \leq x] = \int_{-\infty}^{x} f(u)du$ *for* $-\infty < x < \infty$.

where f(u) is the value of the p.d.f. of x at u.

Definition 5. If f(x) and F(x) are the values of the p.d.f. and the c.d.f. of X at x, then $P[a \leq X \leq b] = F(b) - F(a)$.

For any real constants a and b with $a \leq b$ and $f(x) = \frac{d}{dx} F(x)$, provided the derivative exists.

It is to be noted that $P[X = c] = P[c \leq X \leq c]$

$\int_{c}^{c} f(x)dx = 0$ when X is a continuous random variable.

Example 4. Find the distribution function of the following p.m.f. and represent it graphically.

$p(x) = 1/8$ for $x = 0$
$= 1/4$ for $x = 1$
$= 3/8$ for $x = 2$
$= 1/4$ for $x = 3$
$= 0$ elsewhere.

Solution : By definition, the distribution function F(X) of the random variable X with the p.m.f. p(x) is given by

$$F(x) = \sum_{u \le x} p(u)$$

$$F(-1) = \sum_{u \le -1} p(u) = 0$$

Hence, $F(0) = \sum_{u \le 0} p(u) = 0 + \frac{1}{8} = \frac{1}{8}$

$$F(1) = \sum_{u \le 1} p(u) = 0 + \frac{1}{8} + \frac{1}{4} = \frac{3}{8}$$

$$F(2) = \sum_{u \le 2} p(u) = 0 + \frac{1}{8} + \frac{1}{4} + \frac{3}{8} = \frac{6}{8}$$

$$F(3) = \sum_{u \le 3} p(u) = 0 + \frac{1}{8} + \frac{1}{4} + \frac{3}{8} + \frac{1}{4} = 1$$

$F(x) = 1$ for $x \geq 3$.

Now F(x) can be written in a better way as

$F(x) = 0$ when $-\infty < x < 0$

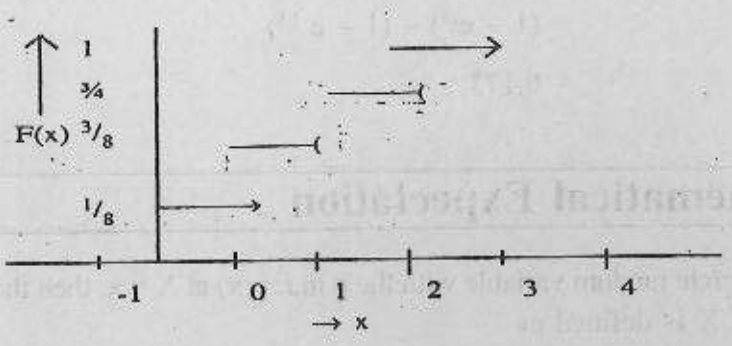
$= \frac{1}{8} \; 0 \le x < 1$

$= \frac{3}{8} \; 1 \le x < 2$

$= \frac{3}{4} \; 2 \le x < 3$

$= 1 \quad 3 \le x < \infty$

The graphical representation of F(x), which is a step function, is as wollows :



That is, at x = 0 there is a jump of 1/8, then at x = 1 an additional jump of $\frac{2}{8}$ and so on.

Example 5. Find the c.d.f. for the following density function.

$f(x) = 3e^{-3x}$, for $x > 0$

$= 0.$ elsewhere.

Hence find P $[0.5 \leq x \leq 1]$

Solution : For $x > 0$

$$F(x) = \int_{-\infty}^{x} f(u)du = \int_{0}^{x} 3e^{-3u}du$$

$$= -e^{-3u} \Big|_0^x = 1 - e^{-3x}$$

For $x \leq 0$, $F(x) = 0$

Therefore, $F(x) = 0$ if $x \leq 0$

$= 1 - e^{-3x}$ if $x > 0$

Now to evaluate the probability

P $[0.5 \leq x \leq 1]$, we use the definition of c.d.f.

That is,

$P[0.5 \leq x \leq 1] = F(1) - F(0.5)$

$= (1 - e^{-3}) - (1 - e^{-1.5})$

$= 0.173$

2.5 Mathematical Expectation

If X is a discrete random variable with the p.m.f. p(x) at X = x, then the mathematical expectation of X is defined as

$$E(X) = \sum_{x=-\infty}^{\infty} x\, p\,(x).$$

Correspondingly, if X is a continuous random variable with the p.d.f. f(x) at X = x, then the expected value of X denoted by E(X) defined as,

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

That is,

$$E(X) = \sum_{x=-\infty}^{\infty} x\, p\,(x), \text{ if X is a discrete random variable.}$$

$$= \int_{-\infty}^{\infty} xf(x)dx, \text{ if X is a continuous random variable.}$$

In gerneal, if $\phi(x)$ is a function of X which is, again, a random variable, then the mathematical expectation of $\phi(x)$ is defined as,

$$E[\phi(X)] = \sum_{x=-\infty}^{\infty} \phi(x)\, p(x), \text{ if X is discrete.}$$

$$= \int_{-\infty}^{\infty} \phi(x) f(x)dx, \text{ if X is continuous.}$$

(i) If X takes only positive integral values then $E(X) = \sum_{k=1}^{\infty} P(X \geq K)$

(ii) If F(x) be the c.d.f. of X with $X \geq 0$, then $E(X) = \int_{0}^{\infty} [1 - F(x)]\, dx$

Example 6. A discrete random variable X has the following probability mass function. Find E(X).

X	:	1	2	3	4	5
P(x)	:	2/5	2/15	1/15	1/10	1/10

Solution : $E(X) = \sum_{x=1}^{5} xp(x)$

$$= 1 \times \frac{2}{5} + 2 \times \frac{2}{15} + ... + 5 \times \frac{1}{10}$$

$$= \frac{60+16+2+9+6}{30} = \frac{93}{30} = 3.1$$

Example 7. The p.d.f. of lifetime (X) in years of a musical C.D. is given by

$f(x) = C\ (10 - x)$, for $0 \leq x \leq 10$

$= 0$, otherwise

Find the expected life time of the C.D.

Solution : Since f(x) is a p.d.f., so $C \int_0^{10} (10-x)\,dx = 1$ *or*, $C\left[10x - \frac{x^2}{2}\right]_0^{10} = 1$

or $C\ [\ 100 - 50] = 1$ or $C = \frac{1}{50}$.

Therefore, $E(X) = \int_0^{10} xf(x)dx$

$$= C\int_0^{10} (10x - x^2)\,dx$$

$$= C\left[10.\frac{x^2}{2} - \frac{x^3}{3}\right]_0^{10}$$

$$= C\left[500 - \frac{1000}{3}\right] = C.\frac{500}{3} = \frac{1}{50} \times \frac{500}{3} = \frac{10}{3}.$$

So the expected life of the C.D. is 10/3 years i.e. 3 years 4 months.

3.5.1 Physical Interpretation of E(X)

Suppose a discrete random variable X takes n different values $x_1, x_2, ... x_n$ connected with a random experiment. If the random experiment is repeated N times under identical conditions and x_1 occurs f_1 times, x_2 occurs f_2 times and so on then the arithmetic mean of x is

$$\bar{x} = \frac{1}{N}\sum_{1}^{n} x_i f_i$$

$$= \sum_{1}^{n} x_i \left(\frac{f_i}{N}\right)$$

If $N \to \infty$ then from the empirical definition of probability $\underset{N \to \infty}{Lt} \frac{f_i}{N} = p_i$, provided the limit exists.

i.e. X takes the value x_i with probability p_i.

Therefore, as $N \to \infty\ \bar{X} \to E(x)$

Hence the mathematical expectation of a random variable is nothing but limiting form of its arithmetic mean.

If $p_1 = p_2 = \ldots\ldots = p_n = 1/n$, then $E(X) = \frac{1}{n}\sum_{1}^{n} x_i$.

For example, if an unbiased die is thrown N times, the probability of face 1 is $p_1 = 1/16$, face 2 is $p_2 = 1/6$ and so on. In this situation expected value of the uppermost face of a die and the arithmetic mean of the face of a die are exactly same.

i.e. $E(X) = \bar{X} = \frac{1+2+\ldots\ldots+6}{6} = 3.5$.

2.5.2 Some Important Results

i) If x = C, a constant, then E (X) = C

ii) U = CX, C being a constant, then E(U) = C E (X)

iii) E [X - E (X)] = 0

iv) E (X $\pm$ C) = E (X) $\pm$ C, C being a constant.

v) If V = aX + bY, a and b are constant, then E (V) = a E(X) + b E(Y), where X and Y are two random variables defined on the sample space.

vi) If W = XY, where X and Y are independent random variables defined on the same sample space then E (W) = E (X) E (Y).

vii) If X is a random variable and $\phi(x)$ is a function of X, then

$E\{C\ \phi\ (X)\} = CE\{\phi\ (X)\}$.

viii) If $\phi_1(X)$ and $\phi_2\ (X)$ be two functions of X,

then $E[\phi_1(X) + \phi_2(X)] = E[\phi_1(X)] + E[\phi_2(X)]$

Note : Result (v) and (vi) are true for more than two random variables.

2.6 Variance of a Random Variable

Variance is a very important characteristic which measures the dispersion of a random variable. Variance of X denoted by V(X), or σ_x^2 or Var(X) is defined by

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

The positive square root of the variance is called the standard deviation which is denoted by σ_x.

The expression for the variance can be simplified as :

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

$= E[X^2 - 2X\ E(X) + E^2(X)]$, where $E^2(X) = \{E(X)\}^2$

$= E(X^2) - 2\ E(X)\ E(X) + E^2(X)$

$= E(X^2) - 2E^2(X) + E^2(X) = E(X^2) - E^2(X)$.

So, $\sigma_x = \sqrt{E(X^2) - E^2(X)}$

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

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

2.6.1 Some Important Results

(i) If X = c, a constant, then V(X) = 0 and V(X) = 0 implies that X is a constant with probability one.

(ii) $V(X \pm d) = V(X)$, d being a constant.

(iii) If U = cX, c being a constant, then $V(U) = c^2\ V(X)$ and hence $\sigma_u = |c|\sigma_x$.

(iv) If W = aX + bY, a and b are constants and X and Y are two independent random variables, then, $V(W) = a^2\ V(X) + b^2\ V(Y)$.

(v) $E[(X - d)^2] \geq V(X)$, when d is a constant.

2.7 Moments

Here we will consider various types of moments for a random variable X.

(a) The r-th moment about origin or r-th raw moment denoted by μ_r' and is defined by $\mu_r' = E(X^r)$. When r = 1, we have $\mu_1' = E(X)$, which is just the expected value of the random variable X. In view of the importane of E (X) in Statistics, we dnote E(X) by μ.

(b) The r-th moment about mean or simply the r-th central moment is denoted by μ_r and is denoted by μ_r and is defined by $\mu_r = E[(X) - \mu^r]$.

The second central moment, i.e. μ_2, is of special importance in Statistics because it means the spread or scatter or dispersion of a random variable. μ_2 is called the variance of the random variable X and it is denoted by σ_x^2 or V(X) or Var (X).

(c) The r-th factorial moment is denoted by $\mu_{[r]}$ and is defined as $\mu_{[r]} = E[X(X-1)(X-2) ..... (X - r + 1)]$.

(d) The r-th absolute moment about a constant C is denoted by

$$\mu_{[r]}^* = E\left|X - C\right|^r$$

The raw moments can be obtained from factorial moments by using the relations

$$\mu_1' = \mu_{[1]}$$

$$\mu_2' = \mu_{[2]} + \mu_{[1]}$$

$$\mu_3' = \mu_{[3]} + 3\mu_{[2]} + \mu_{[1]}$$

$$\mu_4' = \mu_{[4]} + 6\mu_{[3]} + 7\mu_{[2]} + \mu_{[1]}$$

Similarly, the central moments can be obtained from the raw moments by using the relations :

$$\mu_1 = \mu_1' - \mu_1' = 0$$

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

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

$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4 .$$

2.8 Moment Generating Function

Though the moments of most of the probability distributions can be evaluated directly, an alternative method is of immense value in statistical literature. This method is moment generating function (m.g.f.)

The moment generating function, $M_x(t)$, of a random variable X, if it exists, is given by $M_x(t) = E(e^{tx})$

$$= \sum_{x=-\infty}^{\infty} e^{tx} p(x) \text{ when X is discrete}$$

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

where t is a real constant and we are usually interested in values of t in the neighbourhood of 0. Expanding e^{tx} as a power series, we get

$$M_x(t) = E\left[1 + \frac{tX}{1!} + \frac{t^2X^2}{2!} + \frac{t^3X^3}{3!} + ... + \frac{t^rX^r}{r!} + ...\right]$$

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

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

where μ_r' is the r-th moment about the origin and is the coefficient of $\frac{t^r}{r!}$ in the power series expansion of $M_x(t)$. It may also be noted that μ_r' will be determined by the r-th derivative of $M_x(t)$ at t = 0 i.e.

$$\mu_r' = \left[\frac{d^r M_x(t)}{dt^r}\right]_{t=0}$$

The m.g.f. does not exist for some probability distributions. But this function, if it exists, completely determines the probability distribution of a random variable. If two random variables have same m.g.f. they must have the same probability distribution.

Some important results relating the m.g.f. are :

If a and b are constants, then

(i) $M_{x+a}(t) = E[e^{t(x+a)}] = e^{at} E(e^{tx}) = e^{at} M_x(t)$

(ii) $M_{bx}(t) = E[e^{tbx}] = e(e^{btx}) = M_x(bt)$

(iii) $M_{a+bx}(t) = E[e^{t(a+bx)}] = e^{at} E(e^{(bt)x}) = e^{at} M_x(bt)$.

To obtain moments about the mean, that is, the central moments, we may use

$M_{x-\mu}(t) = E(e^{t(x-\mu)}) = e^{-t\mu} M_x(t)$.

That is, the r-th central moment, μ_r, is the coefficient of $\frac{t^r}{r!}$ in $M_{x-\mu}(t)$. So μ_1 will be obtained by computing the r-th derivative of $M_{x-\mu}(t)$ at $t = 0$.

i.e., $$\mu_r = \left[\frac{d^r}{d\mu^r} M_{x-\mu}(t)\right]_{t=0}.$$

Note : $E(t^x)$ generates the factorial moments in the sense that

$$\mu_{[r]} = \left[\frac{d^r E(t^x)}{dt^r}\right]_{t=1}.$$

2.9 Characteristic Function

The moment generating function does not exist for all probability distributions. However, another function which always exists is called the characteristic function (C.F.). It is defined by $E(e^{itX})$ with $i = \sqrt{-1}$ for all real t. If the characteristic function of a random variable X is denoted by $\phi_X(t)$, then

$$\phi_x(t) = E(e^{itx}).$$

Two probability functions are identical if their characteristic functions are identical.

Example 8. Consider a case where an unbiased coin is tossed 3 times and is X is the random variable denoting number of heads occurring, compute V(X) and σ_x.

Solution : From Example 2 the probability distribution of X is

x	:	0	1	2	3	Total
p(x)	:	1/8	3/8	3/8	1/8	1

Now $V(X) = E(X^2) - E^2(X)$.

$$E(X) = \sum_{x=0}^{3} xp(x)$$

$$= 0\times\frac{1}{8}+1\times\frac{3}{8}+2\times\frac{3}{15}+3\times\frac{1}{8}$$

$$= 0+\frac{3}{8}+\frac{6}{8}+\frac{3}{8}+\frac{3}{8}=\frac{12}{8}=1.5$$

$$E(X^2) = \sum_{x=0}^{3} x^2p(x)$$

$$= 0^2\times\frac{1}{8}+1^2\times\frac{3}{8}+2^2\times\frac{3}{15}+3^2\times\frac{1}{8}$$

$$= 0+\frac{3}{8}+\frac{12}{8}+\frac{9}{8}=\frac{24}{8}=3$$

$$\therefore V(X) = E(X^2) - \{E(X)\}^2$$

$$= 3 - (1.5)^2 = 0.75$$

S.D. $= \sigma_x = \pm\sqrt{V(X)} = \pm\sqrt{0.75} = +0.86\ (appr.)$

Example 9. Find the standard deviation of a random variable X that has the following probability density function.

$$f(x) = \frac{x}{2} \quad \text{for } 0 < x < 2$$

$$= 0 \quad \text{elsewhere}$$

Solution : First of all, we calculate V(X)

$$V(X) = E(X^2) - \{E(X)\}^2$$

$$\text{Now } E(X) = \int_{-\infty}^{\infty} xf(x)dx = \int_{0}^{2} x.\frac{x}{2}dx$$

$$= \frac{1}{2}\left[\frac{x^3}{3}\right]_0^2 = \frac{8}{6} = \frac{4}{3}.$$

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)dx \int_0^2 x^2 . \frac{x}{2} dx$$

$$= \frac{1}{2}\int_0^2 x^3 dx = \frac{1}{2}\left[\frac{x^4}{4}\right]_0^2 = \frac{1}{2} \cdot \frac{16}{4} = 2.$$

So, $V(X) = 2 - \left(\frac{4}{3}\right)^2 = \frac{2}{9}$

Hence $\sigma_x = +\sqrt{V(X)} = +\frac{\sqrt{2}}{3}$

2.10 Skewness and Kurtosis

The lack of symmetry of a probability distribution is called its skewness. If a probability distribution is symmetric then it can be checked that the odd order central moments, $m_{2r+1} = 0$, r = 0, 1, 2, ... The distribution can be skewed to the left (positively skewed), skewed to the right (negatively skewed) and symmetric (zero (skewed). The skewness or asymmetry of a probability distribution can be measured by $\gamma_1 = \frac{\mu_3}{\mu_2^{3/2}}$ i.e., the probability distribution is positively skewed, negatively skewed and symmetric, according as $\gamma_1 > 0$, $\gamma_1 <$ and $\gamma_1 = 0$.

Kurtosis refers to the degree of peakedness of a probability distribution. Kurtosis or peakedness of a probability distribution is usually measured by $\gamma_2 = \frac{\mu_4}{\mu_2^2} - 3 = \frac{\mu_4}{\sigma^4} - 3$

The distribution for which $\gamma_2 > 0$, $\gamma_2 < 0$ and $\gamma_2 = 0$ are called leptokurtic, platykurtic and mesokurtic respectively.

2.11 Median and Mode of a Random Variable

The median may be defined in terms of the distribution function of a random variable. The median (M_e) value is such that the probability for X to be less than or equal to M_e and the probability for X to be greater than or equal to it are both equal.

In discrete case M_e is such that

$P(X < M_e) < 1/2 \leq P(X \leq M_e) = F(M_e)$.

In the continuous case M_e is such that

$$\int_{-\infty}^{Me} f(x)dx = \int_{Me}^{\infty} f(x)dx = \frac{1}{2}.$$

The mode (M_o) of a random variable X is that value of X which has the highest probability. So the mode of a discrete random variable is the most probable value of X. In continuous case, M_o is such that the p.d.f. f(x) is maximum at $X = M_o$. That is, $f'(M_o) = 0$ and $f''(M_o) < 0$.

2.12 Mean Deviation

The mean deviation of a random variable X about A (a measure of central tendency) is defined by $MD_A = E(|X - A|)$, $A = E(X)$ or M_e or M_o provided the expectation exists. But the problem with the MD_A is that, if it is defined, it may not be easily amenable to algebraic treatment for its modulus value. It can be verified that for any random variable X, $MD_A \geq MD_{Me}$

2.13 Bivariate Probability Distribution

Let us consider two random variables X and Y. X takes the values $x_1, x_2, \ldots x_m$ and Y takes the values $y_1, y_2, \ldots y_n$. Here p_{ij} notes the probability that (X, Y) takes value (x_i, y_j), that is, $P[X = x_i, Y = y_j] = P_{ij}$, i = 1, 2, ..., m and j = 1, 2, ..., n.

The joint distribution of two rondom variables is shown in the following two-way table.

X	Y y_1 y_2 ..... y_n	Total
x_1	p_{11} p_{12} ..... p_{1n}	p_{10}
x_2	p_{21} p_{22} ..... p_{2n}	p_{20}
.	.	.
.	.	.
x_m	p_{m1} p_{m2} ..... p_{mn}	p_{m0}
Total	p_{01} p_{02} ..... p_{0n}	1

With

$$p_{i0} = P[X = x_i] = \sum_{j=1}^{n} pij$$

$$p_{0j} = P[Y = y_j] = \sum_{i=1}^{m} pij$$

The marginal distribution of X is as follows :

X	:	x_1	x_1	.	.	.	x_m	Total
$P(X=x_i)$	:	p_{10}	p_{20}	.	.	.	p_{m0}	1

Similarly, the marginal distribution of Y is as follows :

Y	:	y_1	y_2	.	.	.	y_n	Total
$P(Y = y_j)$	:	p_{01}	p_{02}	.	.	.	p_{0n}	1

Using the marginal distribution of X one can compute

$$\mu_x = E(X) = \sum_{i=1}^{m} x_i p_{io} \text{ and}$$

$$\sigma_x^2 = V(X) = E[X - E(X)]^2 = E(X^2) - \mu_x^2 = \sum_{1}^{mi} x_i^2 p_{io} - \mu_x^2 .$$

Similarly, the marginal distribution of Y gives.

$$\mu_y = E(Y) = \sum_1^n y_j p_{oj} \text{ and}$$

$$\sigma_y^2 = E[Y - E(Y)]^2 = E(Y^2) - \mu_y^2 = \sum_1^n y_j^2 p_{oj} - \mu_y^2 .$$

Two random variables X and Y are said to be independent if
$P[X = x_i, Y = y_j] = P(X = x_i) P(Y = y_j)$.
i.e. $p_{ij} = p_{i0}p_{0j}$, for i = 1, 2, ..., m and j = 1, 2, ..., n.
The correlation coefficient between X and Y

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}, \quad -1 \le \rho_{xy} \le 1,$$

where

σ_{xy} = cov (x, y) = E [{(x–E(X)}{(Y–E(Y)}]
= E(xy) – $\mu_x\mu_y$.

$$E(XY) = \sum\sum x_i y_j p_{ij}$$

The conditional distribution of $X = x_i$, given that $Y = y_j$ is

$$P(X = x_i / Y = y_j) = \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)} = \frac{p_{ij}}{p_{oj}}.$$

Similarly, the conditional distribution of $Y = y_j$, given that $X = x_i$ is

$$P(Y = y_j / X = x_i) = \frac{P(Y = y_j, X = x_j)}{P(X = x_i)} = \frac{p_{ij}}{p_{io}}.$$

Theorem 1. Let X and Y be two jointly distributed random variables, then

$$E(X + Y) = E(X) + E(Y).$$

Proof : Let Z = X + Y be a random variable that takes values $Z_{ij} = x_i + y_j$ with probability p_{ij}, i = 1, 2, ..., m and j = 1, 2, ..., n.

Therefore, from the definition of expectation

$$E(Z) = \sum\sum z_{ij} p_{ij}$$

$$= \sum_i \sum_j (x_i + y_j) p_{ij} = \sum_i \sum_j x_i p_{ij} + \sum_i \sum_j y_i p_{ij}$$

$$= \sum_i x_i \sum_j p_{ij} + \sum_j y_j \sum_i p_{ij}$$

$$= \sum_i x_i \, p_{io} + \sum y_j p_{oj} = E(X) + E(Y).$$

This theorem can easily be extended to the case of several random variables.

Theorem 2. If X and Y are independent random variables then E(XY) = E(X) E(Y)

Proof : Let V = XY be a random variable which takes values $V_{ij} = X_i Y_j$ with probability p_{ij}, i = 1, 2, ..., m and j = 1, 2, ..., n.

So, by the definition of expectation,

$$E(XY) = \sum_i \sum_j V_{ij} p_{ij} = \sum_i \sum_j x_i y_j p_{ij}$$

$$= \sum_i \sum_j x_i y_j p_{io} p_{oj} = \sum_i x_{i \, p_{oi}} \sum_j y_j p_{oj} = \text{ E(X) E(Y).}$$

This resullt can also be extended to more than two mutually independent random variables.

Theorem 3. If X and Y are independent random variables then $\rho_{xy} = 0$.

Proof : From Theorem 2

$$\sigma_{xy} = \text{Cov(X, Y)} = \text{E(XY)} - \text{E(X) (Y)} = 0.$$

Therefore, $\rho_{xy} = \dfrac{\sigma_{xy}}{\sigma_x \sigma_y} = \dfrac{0}{\sigma_x \sigma_y} = 0$.

Note : If two random variables are independent, then they must be uncorrelated (i.e. $\rho_{xy} = 0$) but the converse is not generally true. If each of the random variable takes two distinct values only, then the converse is true.

Theorem 4. If X and Y are two jointly distributed random variables, then

$$V(aX + bY) = a^2 V(X) + b^2(Y) + 2ab \text{ Cov } (X, Y),$$

where a and b are constants.

Proof : Let Z = aX + bY

So E(Z) = E(aX + bY) = aE (X) + bE(Y).

Now,

$$V(aX + bY) = E[Z - E(Z)]^2$$

$$= E[aX + bY - aE(X) - bE(Y)]^2$$

$$= E[a\{X - E(X)\}]^2 + E[b\{Y-E(Y)\}]^2 + 2E[ab\{(X-E(X)\}\ \{(Y-E(Y)\}]$$

$$= a^2E[X-E(X)]^2 + 2^2E[Y-E(Y)]^2 + 2ab\ E[\{X-E(X)\}\{Y-E(Y)\}]$$

$$= a^2\ V(X) + b^2\ V(Y) + 2ab\ Cov\ (X, Y).$$

Corollary : If X and Y are uncorrelated then $V(aX + bY) = a^2\ V(X) + b^2V(Y)$.

In particular, $V\ (X+Y) = V(X) + V(Y) = V(X-Y)$

Theorem 4 can be extended to more than two random variables. For k random variables $X_1, X_2, \ldots X_k$ the general result is

$$V\left(\sum_{i=1}^{k} a_i X_i\right) = \sum_{i=1}^{k} a_i^2\ V(X_i) + 2\sum_{i=1}^{k}\sum_{j=1}^{k} a_i b_j\ Cov\ (X_i, Y_j) \text{ with } i < j.$$

Example 10 : X and Y are two randam variables having the joint probability distribution as given below :

X \ Y	1	3	5
2	.05	.10	.25
4	.15	.05	.15
6	.10	.10	.05

(i) Give the marginal distribution of X and marginal distribution of Y.

(ii) Compute P (X + Y > 8). P (X + Y= 7), P (Y = 3 / X = 6) and P (X = 4 / Y = 5)

(iii) Compute the correlation coefficient between X and Y.

(iv) Are X and Y independent?

Solution :

The joint distribution is as follows :

X \ Y	1	3	5	Margianal Total
2	0.05	0.10	0.25	0.40
4	0.15	0.05	0.15	0.35
6	0.10	0.10	0.05	0.25
Marginal Total	0.30	0.25	0.45	1

(i) Marginal distribution of X :

x :	2	4	6	Total
p(x) :	0.40	0.35	0.25	1

Marginal distribution of Y :

y :	1	3	5	Total
p(y) :	0.30	0.25	0.45	1

(ii) $P[X + Y > 8]$

$= P[X = 4, Y = 5] + P[X = 6, Y = 3] + P[X = 6, Y = 5]$

$= 0.15 + 0.10 + 0.05 = 0.30.$

$P[X + Y = 7] = P[X = 2, Y = 5] + P[X = 4, Y = 3] + P[X = 6, Y = 1]$

$= 0.25 + 0.05 + 0.10 = 0.40.$

$P[Y = 3 / X = 6] = P[Y = 3, X = 6] / P[X = 6]$

$= 0.10 / 025 = 0.40.$

$P[X = 4/Y = 5] = P[X = 4, Y = 5] / P[Y = 5]$

$= 0.15 / 0.45 = 1/3.$

(iii) Here, from the marginal distribution of X

$\mu_x = E(X) = 2 \times 0.40 + 4 \times 0.35 + 6 \times 0.25 = 0.8 + 1.4 + 1.5 = 3.7$

and

$\sigma_x^2 = V(X) = E(X^2) - \mu_x^2$

Now $E(X^2) = 2^2 \times 0.40 + 4^2 \times 0.35 + 6^2 \times 0.25 = 1.6 + 5.6 + 9.0 = 16.2$

So $\sigma_x^2 = E(X^2) - \mu_x^2 = 16.2 - (3.7)^2 = 2.51$ and $\sigma_x = +1.584$.

Similarly, from the marginal distribution of Y

$\mu_y = E(Y) = 1 \times 0.30 + 3 \times 0.25 + 5 \times 0.45 = 0.03 + 0.75 + 2.25 = 3.3$

and

$\sigma_y^2 = V(Y) = E(Y^2) - \mu_y^2$.

Now $E(Y^2) = 1^2 \times 0.30 + 3^2 \times 0.25 + 5^2 \times 0.45 = 0.3 + 0.75 + 11.25 = 13.8$

So $\sigma_y^2 = E(Y^2) - \mu_y^2 = 13.8 - (3.3)^2 = 2.91$ and $\sigma_y = +1.7059$

$\sigma_{xy} = Cov(X, Y) = E(XY) - E(X)\,E(Y)$

$E(XY) = 2 \times 1 \times 0.05) + (4 \times 1 \times 0.15) + (6 \times 1 \times 0.10) + (2 \times 3 \times 0.10) + (4 \times 3 \times 0.05) + (6 \times 3 \times 0.10) + (2 \times 5 \times 0.25) + (4 \times 5 \times 0.15) + 6 \times 5 \times 0.05) = 11.3$

So $\sigma_{xy} = E(XY) - E(X)\,E(Y) = 11.3 - 3.7 \times 3.3 = -0.91$

Hence, the correlation coefficient between X and Y = $\rho_{xy} = \dfrac{\sigma_{xy}}{\sigma_x \sigma_y} = -0.3368$.

(iv) X and Y are independent if $P(X = x, Y = y) = P(X = x)\,P(Y = y)$

Here $0.5 = P(X = 2, Y = 1) \neq P(X = 2)\,P(Y = 1) = 0.30 \times 0.40 = 0.12$

So, X and Y are not independent.

Definition 6 : For a pair of discrete random variables X and Y, a function $p(x, y) = P(X = x, Y = y)$ is called the joint probability mass function of X and Y, if p(x, y) satisfies the conditions :

(i) $p(x, y) \geq 0$, for each pair of values of (x, y)

(ii) $\sum_x \sum_y p(x,y) = 1$, where the summation is over all possible pairs (x, y).

Definition 7 : For a pair of continuous random variables X and Y, a function f(x,y), such that $\int_{c}^{d}\int_{a}^{b} f(x,y)dxdy = P[a \leq X \leq b,\ c \leq Y \leq d]$ is called the joint probability density function of X and Y if f(x, y) satisfies the conditions :

(i) $f(x, y) \geq 0$, for all x and y

(ii) $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = 1$

2.14 Exerise

1. Define a random variable. How do you distinguish between the discrete and continuous random variables? Illustrate with examples.
2. What do you mean by probability mass function and probability density function? Give examples in each case.
3. How do you define mode and quartiles of a probability distribution?
4. What is meant by mathematical expectation of a random variable? State its important properties. Give an example to illustrate its usefulness.
5. Define an m.g.f. How do you use this to compute different moments? State its important properties.
6. Compute the measures of skewness and kurtosis from the following p.m.f.

x	:	1	2	3	4	5	6
p(x)	:	0.05	0.15	0.20	0.45	0.10	0.05

7. If the probability that the price of a stock will remain the same is 0.36, the probability that its value will increase by Rs. 1.00 or Rs. 2.00 per share are 1.00 per share is 0.14, compute the expected gain per share.

8. Let (x, y) be a pair of discrete random variable with the joint distribution

X \ Y	1	2	3
1	5/27	4/27	2/27
2	1/27	3/27	3/27
3	3/27	4/27	2/27

a. Obtain the marginal probobility distribution of x and y.

b. Compute $P(X > Y)$, $P(X = Y)$, $P(2Y + X > 5)$, $P(X = 2/Y = 3)$ and $P(Y = 1/ X = 2)$.

c. Calculate the correlation coefficient.

d. Are X and Y independent?

9. Let two random variables X and Y be such that $E(X) + E(Y) = 0$, $V(X) - V(Y) = 0$ and $1 + p_{xy} = 0$. What is the relationship between X and Y?

10. For two random variables X and Y

$E(X) = 50$, $E(Y) = 60$, $V(X) = 160$, $V(Y) = 360$.

$\rho_{xy} = 0.75$, Compute

(a) $COV(X + Y, Y)$

(b) $V(5X - 3Y)$

(c) Correlation coeffcient between $(5X - 3Y)$ and $(5X + 3Y)$.

11. Prove that two uncorrelated random variables are independent, if each of the variables takes two distint values only.

12. Find K if the following function is the p.m.f. or p.d.f.

(a) $f(x) = \frac{1}{2x}$ for $x = 1, 2, 3, 4$

$= K$ for $x = 5$

$= 0$ elsewhere.

(b) $f(x) = K\theta e^{-\theta x},\ x > 0$

$= 0$ elsewhere

where θ is constant with $\theta > 0$

13. Evaluate the distribution function and compute E(X) and V(X) for the following probability function

(a) $p(x) = 1/2$ for $x = 1$

$= 1/4$ for $x = 2$

$= 1/4$ for $x = 3$

$= 0$ elsewhere.

(b) $f(x) = 4x/5, \quad 0 < x \leq 1,$

$= 2(x - x)/5, \quad 1 < x \leq 2.$

14. Obtain central moments μ_2, μ_3 and μ_4 from the m.g.f. $M(t) = e^{3t}$. Hence compute measures of skewness and kurtosis.

5. A discrete random variable X has the following p.m.f.

x	:	1	2	3	4	5	6	7
p(x)	:	3a	5a	7a	8a	10a	6a	9a

(a) Determine the value of a.

(b) $P[X \leq 4]$ and $P[2 \leq X \leq 5]$

(c) Find the minimum value of K such that

$P[X \leq K] \geq 0.5$.

16. For what value of k, $p(x) = \frac{2x}{k(k+1)}$ for $x = 1, 2, \ldots k$ is the probability distribution?

17. The p.d.f. of a random variable X is given by

$p(x) = 6x(1 - x)$ for $0 < x < 1$

$= 0$ elsewhere

Find $P(X < 0.35)$, $P(X > 0.5)$ and $P(0.25 \leq X \leq 0.75)$.

18. In a certain city the daily consumption of water (in millions of liters) is a random variable with the density function

$f(x)=\frac{1}{9}xe^{-\frac{x}{3}}$ for $x > 0$

$= 0$ elsewhere.

What is the probability that on a given day

(a) the water consumption in the city is not more than 6 milion liters?

(b) the water supply is inadequate if the daily consumption of this city is 9 milion liters?

19. Evaluate the probability mass function from the following distribution function :

$F(x) = 0, \quad x < 1$

$= 1/4, \quad 1 \leq x \leq 2$

$= 1/2, \quad 2 \leq x < 3$

$= 1, \quad x \geq 3.$

20. Evaluate the density function from the following distribution function

$F(x) = 2x^2/5 \quad 0 < x \leq 1$

$= -3/5 + 2(3x - x^2/2)/5, \quad 1 < x \leq 2$

$= 1, \quad x > 2.$

21. A discrete randam variable can take all possible integral values from 1 to k each with probability 1/k. Find the mean and variance of the distribution.

22. For a distribution with the pdf

$f(x) = \frac{2}{9}x, \quad 0 \leq x \leq 3,$

find the mean and the standard deviation.

23. A person plays a game of throwing a dix under the condition that he could get as many rupees as the number of points on the uppermost face. Find the expectation and variance of his winnings.

24. A random variable X has the pdf

$f(x) = ax^2, \ 0 \leq x \leq 1$

0 else where

Find the value of the constant C and hence find

(i) $p[0 \leq X \leq \frac{1}{2}]$, (ii) $p(X > \frac{3}{4})$.

Unit 3 ❑ Discrete Probability Distribution

Structure

3.1 Introduction

In Statistics the main problem is to infer some characteristics of a population or universe. This is done by observed frequency distribution based on a sample by using theoretical distributions. The construction of observed frequency distribution and its various descriptive statistical measures have been done earlier. Here we are interested in studying the population through theoretical probability distribution of some random variables. If a random variable satisfies the condition of a theoretical probability distribution, the distribution can be fitted to the observed data. These distributions are of two types, depending upon the random variable which is discrete or continuous. In this chapter we will consider some probability distribution of discrete random variables, namely Uniform distribution, Binomial distribution, Poisson distribution and Geometric distribution. Also we will discuss some important properties of these distributions.

3.2 Uniform Distribution

It is the simplest of all probability distributions where each possible value of a random variable has equal probability of occurrence. Such a proability distribution is called uniform distribution. If a random variable X takes k different values $x_1, x_2, \ldots\ldots, x_k$ with eqaul probability, then the p.m.f of X is defined as

$p(x) = 1/k$, for $x = x_1, x_2, \ldots\ldots, x_k$ where $x_i \neq x_j$ for $i \neq j$, i 1, 2, ..., k and $j = 1, 2, \ldots\ldots, k$.

By notation $X \sim U(k)$, means that X follows the uniform distribution with parameter k.

In accordance with the definition of the p.m.f. of the random variable X, the mean

$$\mu = E(X) = \sum_{i=1}^{k} x_i(1/k)$$ and the variance

$$V(X) = E(X-\mu)^2 = \sum_{i}^{k} (x_i - \mu)^2 (1/k)$$

In a special case, when $x_i = i$, the uniform distribution becomes

$p(x) = 1/k$ for $x = 1, 2, \ldots\ldots, k$

With this form of uniform distribution,

the mean = $\mu = E(X) = (k + 1)/2$

and the variance = $\sigma^2 = (k^2 - 1)/12$

This form can be applied to an example of the number of points appeared if we roll a balanced die.

3.3 The Binomial Distribution

By the word 'trial' we mean an attempt to produce a particular event which is neither certain nor impossible. For example, in tossing a coin to get a head is a trial. When trials are repeated, they form a series of trials. Trials are said to be independent

if the probability of an event is not affected by any other trials.

A series of trials are said to be Bernoullian series if,

(i) any trial results two outcomes : a success or a failure.

(ii) the trials are independent,

(iii) the probability of success at any trial is constant (p).

Let X be the random variable denoting the number of successes in n independent trials. Here X can take the values 0, 1, 2, ..., n with non-zero probabilities. If there are exactly x successes, then the remaining (n-x) are failures. Let the probability of success be p and that of failure be $q = 1 - p$. The probability of getting x successes (S) and consequently (n–x) failures (F) in n independent trials in a specified order (say) SSFSFS ..... SFS is given by the expression.

$$P[SSFSFS \ldots\ldots\ldots SFS] = P(S)\ P(S)\ P(F)\ P(S) \ldots\ldots\ P(F)\ P(S)$$

$$= p.p.q.p. \ldots\ldots\ q.p.$$

$$= p^x\ q^{n-x}$$

Now x successes out of n trials can occur in ${}^nC_x$ ways and the probability of each of these ways is p^xq^{n-x}. Hence the probability of exactly x successes in n Bernoullian trials is given by

$$p(x) = {}^nC_xp^xq^{n-x} \quad x = 0, 1, 2, \ldots\ldots\ldots, n$$

$$= 0 \qquad \text{elsewhere}$$

Here $p(x) \geq 0$, for all x and

$$\sum_{i=0}^{n} p(x) = \sum_{i=0}^{n} {}^nC_xp^xq^{n-x} = (q+p)^n = 1.$$

So this is the p.m.f of the binomial distribution. By notation $X \sim$ bin (n, p), means that X follows the binomial distribution with parameters n and p.

This distribution is called the binomial distribution because p(x) is the $(x+1)^{th}$ term in the binomial expansion of $(q+p)^n$.

3.3.1 Important Properties of the Binomial Distribution

1. Binomial distribution is a probability mass function of a discrete random variable X which takes finite unmber of values 0, 1, 2, ..., n. The distribution is

completely defined by the two parameters n and p.

2. Mean of the distribution = $E(X) = np$

Proof : Mean = $E(X)$

$$= \sum_{x=0}^{n} x\, p(x)$$

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

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

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

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

$$= np \sum_{y=0}^{n-1} (n-1)!/\{y!(n-y-1)!\}\, p^y q^{n-y-1}$$, where $y = x - 1$ when $x = 1$, $y = 0$ and when $x = n$, $y = n - 1$

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

$$= np$$

3. Variance = $V(X) = npq$ and S.D. $= +\sqrt{V(X)} = +\sqrt{npq}$

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

$= E(X^2) - E^2(X)$

$= E[X(X-1)] + E(X) - E^2(X)$.

Now, $E[X(X-1)] = \sum_{x=2}^{n} x(x-1) \; {}^{n}C_x p^x q^{n-x}$

$$= \sum_{x=2}^{n} n!/\{(x-2)!\,(n-x)!\}\, p^x q^{n-x}$$

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

$$= n(n-1)p^2 \sum_{0}^{n-2} (n-2)!/\{y!(n-y-2)!\}\, p^y q^{n-y-2}$$

$$= n(n-1)\, p^2\, (q+p)^{n-2}$$ where $x-2=y$, $x=y+2$

$$= n(n-1)p^2$$ when $x=2$, $y=0$; when $x=n$, $y=n-2$

So, $V(X) = n(n-1)\, p^2 + np - n^2p^2$

$$= np\,(1-p)$$

$$= npq$$

4. Third and fourth order central moments are

$\mu_3 = npq\,(q-p)$ and $\mu_4 = 3n^2p^2q^2 + npq\,(1-6pq)$.

5. Skewness $\gamma_1 = \dfrac{p-q}{\sqrt{npq}}$. So the distribution is symmetric if $p = q = 1/2$.

6. Kurtosis $(\gamma_2) = (1 - 6pq)/npq$.

7. The recursion relation for the probabilities is as follows :

$$p(x) = \frac{(n-x+1)}{xq}\, p(x-1)$$

8. Binomial distribution tends to Normal distribution as n is large enough.

9. The moment generating function is

$M_x(t) = E(e^{tx}) = (q+pe^t)^n$.

10. A recursion relation concerning three consecutive central moments is

$$\mu_{r+1} = pq\left[nr\mu_{r-1} + \frac{d\mu_r}{dp} \right],\ r = 1, 2, 3....$$

11. The distribution may be unimodal as well as bimodal,

Mode $= M_o = [n+1)p]$, if $(n+1)p$ is not an integer

$= (n+1)p,\ (n+1)p - 1$, if $(n+1)p$ is an integer

where $[m]$ = greatest integer contained in m.

12. If X and Y are two independent binomial variates with parameters (n_1, p) and (n_2, p) respectively then $(X + Y)$ is also a binomial variate with parameter as

$(n_1 + n_2, p)$. The result holds for more than two binomail variables, provided p remains the same.

13. This distribution can be obtained as a limiting form of hypergeometric distribution.

14. Cumulative probability function :

$$P[X \leq x] = F_x(k) = \sum_{x=0}^{k} p(x) = I_q(n-k,\ k+1),$$

where $I_q(s,t) = \int_0^q u^{s-1}(1-u)^{t-1}\,du \Big/ \int_0^1 u^{s-1}(1-u)^{t-1}\,du$.

The function I_q (s, t) is extensively tabulated by Karl Pearson in tables of Incomplete Beta Function.

15. The r-th order factorial moment,

$$\mu_{[r]} = E[x(x-1) \ldots\ldots (x-r+1)] = {}^nP_r\, p^r, \quad r = 1, 2, 3, \ldots\ldots, n.$$

16. Mean deviation about mean of a binomial variate is

$$MD_{np} = E \mid X - np \mid = 2mq\ {}^nC_m p^m q^{n-m}$$

$$= \sqrt{(2npq/\pi)}, \text{ when n is large,}$$

where m is the largest integer contained in (n+1)p.

3.3.2 Some Real Life Examples

The following are some real life examples of a binomial variable :

(i) Occurrence of heads or tails when a number of times a coin is tossed.

(ii) Occurrence of odd points or even points when a number of times a die is thrown.

(iii) Arrival or non-arrival of ships in a port.

(iv) Infected or non-infected by diseases.

(v) Defective and non-defective items in a lost

(vi) Success and failure in an examination.

(vii) Vegetarian and non-vegetarian in a given populaton.

(viii) Literate and illiterate persons in a community.

Some Problems on Binomial Distribution

1. Show that for the binomial distribution var (x) $\leq \frac{n}{4}$.

Solution : We know that for the binomial distribution Var (x) = npq. So to prove that Var (x) $\leq \frac{n}{4}$ we are to prove that npq $\leq \frac{n}{4}$.

that is, we are to prove that pq $\leq \frac{1}{4}$.

that is, we are to prove that $4pq \leq 1$

For this distribution $p+q = 1$

That is, $p^2 + q^2 + 2pq = 1$

That is, $p^2 + q^2 - 2pq + 4pq = 1$

That is, $(p-q)^2 + 4pq = 1$

That is, $4pq \leq 1$.

$\therefore pq \leq \frac{1}{4}$

That is, $npq \leq \frac{n}{4}$.

The equality sign will hold good when $p=q=\frac{1}{2}$.

2. Find the maximum value of variance of the binomial distribution.

Soluton : For the binomial distribution

$Var(x) = npq = np(1-p) = np - np^2$

To maximise var(x) we proceed as

$$\frac{d\,var(x)}{dp} = n - 2np$$

So, $\frac{var(x)}{dp} = 0$ means $p = \frac{1}{2}$.

Again, $\frac{d^2\,var(x)}{dp^2} = -2n < 0$

Hence, when $p = \frac{1}{2} = q$ the var(x) of the binomial distribution is maximum.In this case var(x) = npq = $n\frac{1}{2}\cdot\frac{1}{2} = \frac{n}{4}$.

3. Let x be a binomially distributed random variable with mean 2 and s.d. $2/\sqrt{3}$. Find the corresponding probability function.

Solution : Here $\mu = 2$ and $\sigma = 2/\sqrt{3}$. That is, $np = 2$ and $\sqrt{npq} = 2/\sqrt{3}$ so that $npq = \frac{4}{3}$.

Here $\frac{npq}{np} = q = \frac{2}{3}$ and $p = \frac{1}{3}$

Since $np = 2$ and $p = \frac{1}{3}$, $n = 6$.

Thus, the probability function of the binomial distribution is

$$f(x) = {}^{6}c_{x} \left(\frac{1}{3}\right)^{x} \left(\frac{2}{3}\right)^{6-x}, \; x = 0, 1, 2, \ldots\ldots 6$$

3.4 Poisson Distribution

Here the Poisson probability distribution will be derived as a limiting case of the binomial distribution. This is a probability distribution of a discrete random variable which assumes a countable infinite number of values. The distribution is due to S.D. Poisson, a French Mathematician.

First let us consider the limiting form of the binomial distribution. Let

i) The number of trials be very large i.e., $n \to \infty$

ii) The probability of success p be very small i.e., $p \to 0$.

iii) $np = \lambda$, a finite quantity, when $n \to \infty$ and $p \to 0$.

Under the above conditions the binomial distribution will be reduced to the poisson distribution in the following way :

$$\lim_{\substack{n \to \infty \\ p \to 0 \\ np = \lambda}} p(x) = \lim {}^{n}C_{x} p^{x} q^{n-x}$$

$$= \lim n! / \{x! (n - x)!\} \; p^{x} q^{n-x}$$

$= \lim n(n-1) \ldots\ldots (n-x+1) / x!\, p^x (1-p)^{n-x}$

$= 1/x!\ \lim (1-1/n)(1-2/n) \ldots\ldots (1-(x-1)/n)\,(np)^x (1-np/n)^{n-x}$, dividing each multiplier by n.

$= 1/x!\ \lim (1-1/n)(1-2/n) \ldots\ldots (1-(x-1)/n)\,\lambda^x (1-\lambda/n)^{n-x}$

$= \lambda^x/x! \lim\{1(1-1/n)(1-2/n) \ldots\ldots (1-(x-1)/n)\} \lim (1-\lambda/n)^{n-x}$

$= e^{-\lambda}\,\lambda^x / x!$

[Since $\lim_{n\to\infty} \{1(1-1/n)(1-2/n) \ldots\ldots (1-(x-1)/n)\} = 1$

$\lim_{x\to\infty} (1-\lambda/n)^n = e^{-\lambda}$ and

$\lim_{x\to\infty} (1-\lambda/n)^x = 1$]

Hence the probability mass function p(x) is reduced to the form

$p(x) = e^{-\lambda}\,\lambda^x / x!$, $x = 0, 1, 2 \ldots\ldots \infty$

Hence the Poisson distribution is defined by the p.m.f

$$p(x) = e^{-\lambda}\lambda^x / x! \quad . \ x = 0, 1, 2 \ldots\ldots \infty$$

$$= 0 \qquad \text{elsewhere.}$$

where $\lambda(>0)$ is the only parameter of the distribution.

By notation $X \sim P(\lambda)$, means that X follows the Poisson distribution with parameter λ.

It is clear that $p(x) \geq 0$, for all x and

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

Thus the pmf of the poisson distribution holds both the properties of being the pmf.

3.4.1 Important Properties of the Poisson Distribution

1. Poisson distribution is a theoretical distribution of a discrete random variable, X which takes infinite number of values 0, 1, 2, ... ∞. The distribution is completely defined by the parameter λ.
2. Mean of the distribution = $E(X) = \lambda$.

Proof. Mean = $E(X) = \sum_{x=0}^{\infty} xp(x)$

$= \sum_{x=1}^{\infty} xe^{-\lambda}\lambda^x / x!$

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

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

$= \lambda e^{-\lambda}\sum_{y=0}^{\infty}\lambda^y / y!$ where $y = x-1$

$\because$ where $x = 1, y = 0$

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

$= \lambda e^0$

$= \lambda$

3. Variance = $V(X) = \lambda$.

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

$= E(X^2) - E^2(X)$

$= E[X(X-1)] + E(X) - E^2(X)$ $\because x^2 = x(x-1)+x$

Now $E[X(X-1)] = \sum_{x=0}^{\infty} x(x-1)e^{-\lambda}\lambda^x / x!$

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

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

$= \lambda^2 e^{-\lambda} \sum_{z=0}^{\infty} \lambda^z / z!$ where $z = x - 2$ when $x = 2, z = 0$.

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

$= \lambda^2$

$V(X) = \lambda^2 + \lambda - \lambda^2 = \lambda$

4. The third and fourth order central moments are $m_3 = \lambda$ and $m_4 = \lambda + 3\lambda^2$.
5. Skewness $(\gamma_1) = 1/\sqrt{\lambda}$, the distribution is always positively skewed.
6. Kurtosis $(\gamma_2) = 1/\lambda$, the distribution is always leptokurtic.
7. The recursion relation for the probabilities is as follows :

   $p(x) = (1 / x)\, p(x - 1)$
8. The moment generating function is

   $M_x(t) = E(x^{tx}) = e^{\lambda(e^t - 1)}$
9. A recursion relation concerning three consecutive central moments is

$$\mu_{r+1} = \lambda\left[r\,\mu_{r-1} + \frac{d\mu_r}{d\lambda} \right], \quad r = 1, 2, 3.....$$

10. The distribution may be unimodal as well as bimodal.

    Mode $= M_0 = [\,\lambda\,]$ if λ is not an integer

    $= A, A - 1$ if λ is an integer

    where $[\,A\,]$ = greatest integer contained in A.
11. If X and Y are two independent Poisson variates with parameter λ_1 and λ_2 respectively then $(X + Y)$ is also a Poisson variate with parameter $(\lambda_1 + \lambda_2)$.
12. This distribution can be obtained as a limiting form of the binomial distribution when p is small, n is large but np is a finite positive number.
13. Cumulative probability function:

$$P[X \le k] = F_x(k) = \sum_{x=0}^{k} p(x) = 1 - I\left(\frac{\lambda}{\sqrt{k+1}}, k\right)$$

where $I(s, t) = \int_0^z e^{-v} v^p \, dv \Big/ \int_0^\infty e^{-v} v^p \, dv$

with $s = z / \sqrt{(t+1)}$

The function I(s, t) is extensively tabulated by Karl Pearson in tables of Incomplete Gamma Function.

14. The r-th order factorial moment,.

$\mu_{(r)} = E[X(X - 1)(X - r + 1)] = \lambda^r, \quad r = 1, 2, 3, ...$

15. Mean deviation about mean of a Poisson variate is

$MD_\lambda = E | X - A | = 2m \, e^{-\lambda} \lambda^m / m!$

where m is the largest integer contained in A + 1.

16. Poisson distribution tends to the normal distribution with mean λ and variance as λ is large enough.

3.4.2 Some Real Life Examples

The following are some examples of a Poisson variable :

(i) Number of accidents per day in a big city.

(ii) Number of printing mistakes per page in a book.

(iii) Number of deaths from a rare disease per year in a given region.

(iv) Number of cars passing through a road crossing per unit interval of time during a busy period.

(v) Number of defects per unit sheet materials (e.g. Paper, metal sheet, cloth etc.)

(vi) Number of defective items per packet manufactured by a reputed company.

(vii) Number of customer visiting a service centre per hour.

(viii) Number of persons born blind per year in a large region.

(ix) Number of goals scored in football match.

(x) Number bacteria present in a given liquid per unit volume,

3.5 Geometric Distribution

Here the discrete random variable X denotes the number of trials that are needed to get the first success. The probability mass function of X is

$$p(x) = pq^x \quad \text{for } x = 0, 1, 2, \ldots, \infty$$
$$= 0 \quad \text{elsewhere}$$

where $q = 1 - p$ and p is the probability of success in any trial of a binomial distribution. Since the different terms in the distribution are the terms in the geometric series, the distribution is known as a geometric distribution.

By notation X~g(p) means that X follows the geometric distribution with parameter p.

Here $p(x) \geq 0$, for all x and

$$\sum_{x=0}^{\infty} p(x) \quad \sum_{x=0}^{\infty} pq^x = p(1+q+q^2+\ldots\ldots)$$
$$= p/(1-q) = p/p = 1$$

3.5.1 Important Properties of the Geometric Distribution

1. Geometric distribution is a probability mass function of a discrete random variable X which takes countable infinite number of values 0, 1, 2 ..... ∞.
   The distribution is completely defined by the parameter p.
2. Mean of the distribution, that is, $E(X) = q/p$.
3. Variance $= V(X) = q/p^2$.
4. The third and the fourth order central moments are respectively
   $\mu_3 = q(1 + q)/p^3$ and $\mu_4 = (q + 7q^2 + q^3)/p^4$.
5. Skewness $(\gamma_1) = (1 + q)/\sqrt{q}$,
6. Kurtosis $(\gamma_2) = (1 + 4q + q^2)/q$
7. The recursion relation for the probabilities is as follows:
   $p(x) = (1 - p)\, p(x - 1)$.

8. The moment generating function is

$M_x(t) = E(e^{tx}) = p(1 - qe^t)^{-1}$, for $qe^t < 1$

Example 1. The probability that an applicant for a driver's license will pass the road test on any given try is 0.75. What is the probability that an applicant will finally pass the test on the fourth try?

Solution : Let X be the random variable denoting number of failures before the first success. Here x =3 and p = 0.75. Hence using the geometric distribution,

P(the applicant will finally pass the test on the fourth try) = p(3)

$= 0.75(1- 0.75)^3$

$= 0.0117$

Example 2. A firm produces an item of which 0.1% are usually defective, if packed then in boxes each containing 500 items. If a wholesaler purchases 1000 such boxes, how many boxes are expected to be :

(i) Free from defective items ?

(ii) One defective item ?

Solution : Let X be a random variable denoting number of defective items in a box of 500 items. Here X follows the binomial distribution with n = 500 and p = 1/ 1000. Since n is sufficiently large, p is very small but np = 0.5 (finite), the binomial distribution can be approximated by the Poisson distribution with p.m.f

$p(x) = e^{-\lambda} \lambda^x / x!, \quad x = 0, 1, 2, \ldots\ldots \infty.$

Here, $\lambda = (1/1000) \times 500$

$= 0.5$

(i) P(a box which is free from defective items) $= P(X = 0)$

$= p(0)$

$= e^{-0.5} = 0.6065$

Hence the number of boxes which are *free* from defective items is

$1000 \times 0.6065 = 6065$

(ii) P(a box with one defective item) $= P(x = 1)$

$= P(I)$

$= e^{-0.5} \times 0.5$

$= 0.3033$

Hence the number of boxes with one defective item is

$1000 \times 0.3033 = 303.3 \simeq 303.$

Example 3. The distribution of printing mistakes in a book of 200 pages is as follows:

No. of printing mistakes in a page (x) :	0	1	2	3	4	5
No. of pages (f) :	112	63	20	3	1	1

Fit a Poison distribution to the above data.

Solution : Let X be the random variable representing the number of printing mistakes per page.

If the given frequency distribution is approximated by the Poisson distribution, then it has only one parameter λ. λ can be estimated from the observed data by the method of moments. The first order raw moment of the Poisson distribution about zero is λ, while the first moment about zero of the observed distribution is average number of printing mistakes per page

$$= \sum xf / \sum f$$

$$= (0 \times 112 + 1 \times 63 + 2 \times 20 + 3 \times 3 + 4 \times 1 + 5 \times 1) / 200$$

$$= 121 / 200$$

$$= 0.605$$

Hence $\hat{\lambda} = \bar{X}_w = 0.605$

X	f	P(x)	Np(x)
0	112	0.545	109
1	63	0.330	66
2	20	0.100	20
3	3	0.020	4
4	1	0.005	1
5	1	0	0
Total	200	1.00	200

Example 4. The manufacturer of a popular brand of T.V. knows from past experience that the probability of a T.V. set failing to work properly during the warranty period

is 0.04. Find the probability that in a sample of 25 sold T.V. sets selected at random, 6 or less will be failing to work properly during the warranty period. Use both binomial and Poisson distributions to compare the results.

Solution: Let X be the random variable denoting that the number of T.V. sets will be failing to work properly in a sample of size n. Here $p = 0.04$, $q = 0.96$, $n = 25$

So by binomial distribution

$$P(x \geq 6) = \sum_{x=0}^{6} {}^{25}C_x (0.04)^x (0.96)^{25-x}$$

$$= 0.9999$$

and by the Poisson distribution with

$\lambda = np = 25 \times 0.04 = 1$, we have

$$P(x \leq 6) = \sum_{x=0}^{6} e^{-1} 1^x / x! = 0.9999$$

Thus two results are identical.

Example 5. The probability that a bulb will fail before 100 hours is 0.3. Bulbs fail independently. If 15 bulbs are tested for lengths of life, what is the probability that in a sample of 10 bulbs a) exactly 4 will fail, b) at most 2 will fail and c) at least one will fail before 100 hours?

Solution : Here X = number of bulbs failing before 100 hours. X follows the binomial

distribution with $n = 10$ and $p = 0.3$.

i) $P(X = 4) = p(4) = {}^{10}C_4 (0.3)^4 (0.7)^6 = 0.2$

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

$= {}^{10}C_0 (0.7)^{10} + {}^{10}C_1(0.3)(0.7)^9 + {}^{10}C_2(0.3)^2 (0.7)^8 = 0.3828$

iii) $P(X \geq 1] = 1 - P(X = 0) = 1 - p(0) = 1 - {}^{10}C_0(0.7)^{10} = 0.9717$

3.6 Exercises

1. Explain the concept of theoretical distribution with reference to a discrete random variable.
2. Explain briefly the characteristics of the binomial and Poisson distributions.
3. Prove that variance of the binomial distribution can not be greater than its mean. Also show that the vanance can not exceed n/4.
4. Show that the mean and variance of a Poisson distribution are equal.
5. For a binomial distribution with parameters n and p, establish the following relationship :

$$\left[\mu_{r+1} = pq\left(nr\mu_{r-1} + \frac{d\mu_r}{dp}\right)\right],$$

where μ_r is the central moment of order r. Hence find measures of skewness and kurtosis of the binomial distribution.

6. Describe Poisson distribution as a limiting form of the binomial distribution.
7. Give some examples of binomial and Poisson variables from our daily life.
8. Determine the mode of the binomial and the Poisson distributions.
9. For a discrete uniform distribution obtain the mean and variance.
10. Show that binomial distribution is symmetric when $p = \frac{1}{2}$.
11. If the probability of having a male or a female child is both 0.5, find the probability that :
    (a) A family's fourth child is their first son.
    (b) A family's fifth child is their first daughter.
12. When taping a television commercial the probability is 0.3 that certain actor will get his lines straight on any one take. What is the probability that he will get his lines straight for the first time on the sixth take?
13. A discrete random variable X follows the uniform distribution and assumes the values 5, 7, 12. 17, 19, 22, 25. Find the following probabilities: P(X=7), $P(X \leq 17)$ and P(X>12).

14. The incidence of a certain occupational disease is such that on the average 30.% workers suffer from it. If 8 workers are selected at random, find the probability that

a) exactly 2 workers suffer from the disease,

b) not more than 2 workers suffer from the disease.

15. The probability of a salesman achieving his sales quota is 0.3.Find the probability that in a random sample of 7 salesmen

a) at least two, b) at most three and c) exactly four will achieve their respective sales quota.

16. A certain factory is turning out with optical lenses. There is a small chance 1/500 for any one lens to be defective. The lenses are supplied in packets of 10. Use the poisson distribution to calculate the approximate number of packets containing no defective lenses in a consignment of 20,000 packets.

17. A period of 100 days was observed for the number of accidents taking place per day in a busy city. The distribution of the observed days according to the number of accidents per day is given below. Assuming Poisson distribution, find the expected frequency.

No. fo accidents :	0	1	2	3	4	5	More than 5
No. of days :	40	23	14	10	7	4	2

18. The screws manufactured by a certain machine were checked by examining samples of 8 screws. The following freqency distribtion gives 200 samples according to number of defective screws they contain. Fit a binomial distribtion to the given data.

Defective screws :	0	1	2	3	4	5	6	7	8
No. of samples :	2	10	24	38	48	35	25	12	6

19. An indstrial area has power brake down one in 30 days, on the avarage. Assuming Poisson distribution, what is the probability that

(i) no power brakes down in the next 7 days?

(ii) more than one power brakes down in the next 7 days?

20. Point out the fallacy in the following statements :

(i) The mean of a binomial distribtion is 9/2 and the standard devition on is $\sqrt{3}$.

(ii) The mean of binominal distribution is 4 and the standard devotion is 3.

21. In a binomial distribution with parameters n and p the mean is 3 and the standard deviation is $\sqrt{2}$. Find the values of n and p and hence find p(x =5).

22. A random variable x follows the poisson distribution with parameter 4. Find the probabilities that x assumes the values (i) 0, 1, 2, and 4; (ii) less than 2 and (iii) at least 3. [It is given that $e^{-4} = 0.0183$]

23. Suppose x has a poisson distribution such that its mean is 2 times its standard deviation. Find ($x \geq 2$).

24. Prove that the binomial distribution is symmetrical if $p = \frac{1}{2}$.

25. For a binomial distribution, the mean and the standard deviation are respectivly 4 and $\sqrt{3}$. Find out the probability of getting a non-nagative value from this distribution.

26. In a shooting competition the probability of a man hitting a target is 0.2. If he fires 5 times, find the probability of hitting the target at least twice.

Unit 4 ❑ Continuous Probability Distribution

Structure

4.0 Introduction

In this chapter we will consider some continuous distributions and discuss their important properties. The distribution will be defined in terms of probability density function (p. d. f.).

4.1 Rectangular Distribution

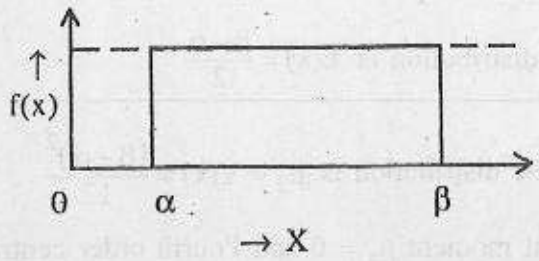
This is also known as the continuous uniform distribution. The distribution has the same probability density at all values through the range of a continuous variable X. The probability density function is

$$f(x) = \frac{1}{\beta - \alpha}, \quad \alpha < x < \beta$$

$$= 0, \text{ elsewhere.}$$

The following figure gives a graphical representation of the distribution. Due to its rectangular shape, it is called a rectangular distribution.

This distribution is also called the uniform distribution because this distribution has uniform or same probability over the range of the values, α to β.



By notation $X \sim R(\alpha, \beta)$ means X follows the rectangular distribution with two paramenters α and β. It is to be noted that $f(x) \geq 0$ and

$$\int_{\alpha}^{\beta} f(x)dx = \int_{\alpha}^{\beta} \frac{1}{\beta-\alpha} dx = 1.$$

So the pdf of the rectangular distribution holds both the properties of being the pdf.

Example : The buses on a certain route run after every 20 minutes. If a person arrives at the bus stop at random, what is the probability that

(a) he has to wait between 5 to 15 minutes,

(b) he gets a bus within 10 minutes

and (c) he has to wait at least 15 minutes?

Solution : Let the random variable X denote the waiting time of the person, which follows a rectangular distribution with p.d.f.

$f(x) = 1/20, \quad 0 \leq x \leq 20.$

Now, (a) $P[5 \leq X \leq 15] = \frac{1}{20}\int_{5}^{15} dx = \frac{15-5}{20} = \frac{1}{2}$

(b) $P[0 \leq X \leq 10] = \frac{10}{20} = \frac{1}{2}$

(c) $P[15 \leq X \leq 20] = \frac{20-15}{20} = \frac{1}{4}$

4.4.1. Some Important Properties

1. This is the p.d.f. of a continuous random variable X. It is completely defined

by two parameters α and β.

2. The mean of the distribution is $E(x) = \frac{\beta - \alpha}{2}$.

3. The variance of the distribution is $\mu_2 = V(x) = \frac{(\beta - \alpha)^2}{12}$.

4. Third order central moment $\mu_3 = 0$ and Fourth order central moment $\mu_4 = \frac{(\beta - \alpha)^4}{80}$.

5. Measure of Skewness $(\gamma_1) = \frac{\mu_3}{\mu_2^{3/2}}$,

   i.e. the distribution is symmetric about $\frac{(\alpha + \beta)}{2}$.

6. Measure of Kurtosis $(\gamma_1) = \frac{\mu_4}{\mu_2^2} - 3$ -1.2, i.e. the distribution is platykurtic.

7. The moment generating function of X is

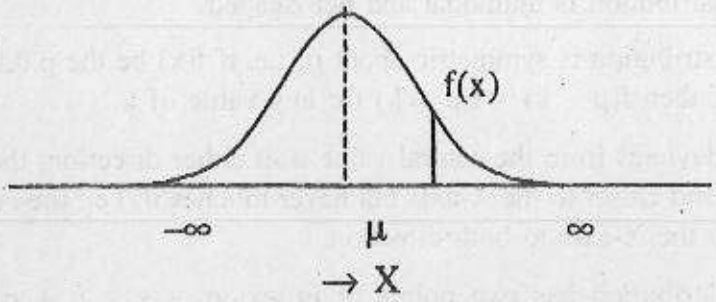
$$M_x(t) = \frac{1}{t(\beta - \alpha)}(e^{t\beta} - e^{t\alpha}), t \neq 0$$

4.2 The Normal Distribution

A continuous random variable X follows the normal distribution and it is referred to as a normal random varable if and only if the probability density function of X is given by $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right\} - \infty < x < \infty$,

where there are two parameters $\mu = E(X)$ and $\sigma^2 = V(X)$. By notation $X \sim N(\mu, \sigma^2)$ and we mean that X follows the normal distribution with two parameters μ and σ^2. It is clear that $f(x) \geq 0$ for all x, and it can easily be verified that $\int_{-\infty}^{\infty} f(x)\,dx = 1$.

The shape of the above distribution has been shown in the following diagram measuring x horizontally and f(x) vertically.



Standard Normal Variable

A variable Z defined by $Z = \frac{X-\mu}{\sigma}$ is called the Standard normal variable.

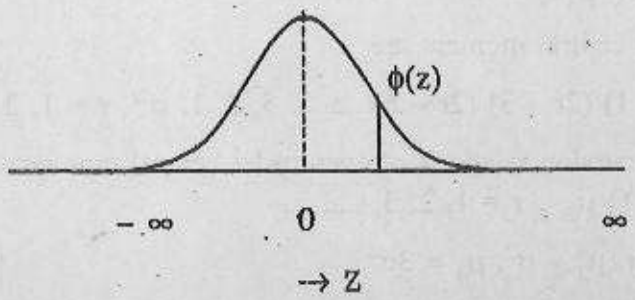
Here $E(Z) = \frac{E(X)-\mu}{\sigma} = \frac{\mu-\mu}{\sigma} = 0$ and $V(Z) = \frac{V(X)}{\sigma^2} = \frac{\sigma^2}{\sigma^2} = 1$.

It can be shown that Z is itself a normal variable with mean zero and unit variance.

The p.d.f. of Z is given by $\phi(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}}, -\infty < z < \infty$

By notation we write $Z \sim N(0, 1)$. That is, Z follows the normal distribution with mean zero and s.d. unity.

The curve representing the standard normal variable has been shown below.



4.2.1 Important Properties of the Normal Distribution

1. The Normal distribution is the p.d.f. of a continuous random variable X, $-\infty < X < \infty$. The distribution is completely defined by two parameters μ and σ^2.
2. Mean, Median and Mode coincide and all are equal to μ, i.e. Mean = Median = Mode = μ.

3. $V(X) = \sigma^2$ and Standard Deviation $= +\sqrt{V(X)} = \sigma$

4. The distribution is unimodal and bell-shaped.

5. The distribution is symmetric about μ, i.e. if f(x) be the p.d.f. of the distribution, then $f(\mu - k) = f(\mu + k)$ for any value of k.

6. As X deviates from the central value μ in either direction, the curve comes closer and closer to the X-axis but never touches it, i.e., the curve is asymptotic to the X-axis to both direction.

7. The distribution has two points of inflexion at $x = \mu \pm \sigma$ which is the solution of the equation $\frac{d^2 f(x)}{dx^2} = 0$. i.e. at these points the normal curve changes its curvature. The curve is convex upwards within the interval $(\mu - \sigma, \mu + \sigma)$ and concave upwards outside this interval.

8. Although a normal variable can theoretically take any value between $-\infty$ and ∞ for all practial purposes it may be assumed to lie between $\mu - 3\sigma$ and $\mu + 3\sigma$. The interval $[\mu - 3\sigma, \mu + 3\sigma]$ is often called the effective range of a normal variable. It may be noted that

   $P[\mu - 3\sigma \leq X \leq \mu + 3\sigma] = 0.9973$ (app.) which is very close to 1.

9. Since the distribution is symmetrical about μ, its odd order central moments vanish, i.e. $\mu_{2r+1} = 0$ for $r = 0, 1, 2, ...$

10. Even order central moment are

    $\mu_{2r} = (2r - 1)(2r - 3)(2r - 5) ..... 5, 3, 1, \sigma^{2r}$, $r = 1, 2, 3, ...$

11. A useful recursion relations of even order central moments is given by $\mu_{2r} = \sigma^2 (2r - 1) \mu_{2r-2}$, $r = 1, 2, 3, ...$

    In particular, $\mu_2 = \sigma^2$, $\mu_4 = 3\sigma^4$.

12. The coefficient of skewness $(\gamma_1) = 0$, that is, the distribution is symmetric.

13. The coefficient of kurtosis $(\gamma_2) = 0$, that is, the distribution is mesokurtic and hence peakedness of the distribution is ideal.

14. The moment generating function of the normal distribution is $M_x(t) = E(e^{tx})$
    $= e^{t\mu + \frac{1}{2}t^2\sigma^2}$

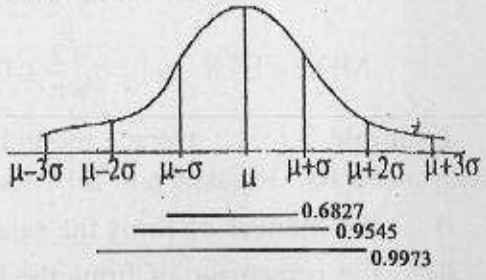
15 The first quartile (Q_1) and the third quartile (Q_3) are equidistant from the mean and $Q_1 = \mu - 0.67\sigma$ (app) and $Q_3 = \mu + 0.67\sigma$ (app). So the quartile deviation $= \frac{Q_3 - Q_1}{2} = 0.67\sigma$ (app).

16. The distribution of probability of a normal curve is as follows :

$P[\mu - \sigma \le X \le \mu + \sigma] = 0.6827$

$P[\mu - 2\sigma \le X \le \mu + 2\sigma] = 0.9545$

$P[\mu - 3\sigma \le X \le \mu + 3\sigma] = 0.9973$



17. The values of c.d.f. of $Z = \frac{X - \mu}{\sigma}$ is $\Phi(k) = P(Z \ge k) = \int \phi(z)\,dz$ are avaliable in statistical tables for different values of k. From the symmetry of the distribution we have $\Phi(k) = \Phi(-k)$ and hence $\Phi(0) = 0.5$.

18. If $X \sim N(\mu, \sigma^2)$, then for any two constants and and b ($b \ge a$) we have

(a) $P[X \ge a] = \int_{-\infty}^{a} f(x)dx = \Phi\left(\frac{a-\mu}{\sigma}\right)$

(b) $P[X \ge b] = \int_{b}^{\infty} f(x)dx = 1 - \Phi\left(\frac{b-a}{\sigma}\right)$

(c) $P[a \le X \le b] = \int_{a}^{b} f(x)dx = \Phi\left(\frac{b-a}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$

19. If X and Y are independent normal variates with means μ_1 and μ_2 and standard deviations σ_1 and σ_2 respectively, then $(aX + bY)$ is also a normal variable with means $a\mu_1 + n\mu_2$ variance $a^2\sigma_1^2 + b^2\sigma_2^2$.

20. Under certain assumptions the binomial distribution and the Poisson distribution can be approximated by normal distribution. i.e. If $X \sim$ bin (n, p) then

$$P[a \le X \le b] \cong \Phi\left(\frac{b + \frac{1}{2} - np}{\sqrt{npq}}\right) - \Phi\left(\frac{a - \frac{1}{2} - np}{\sqrt{npq}}\right)$$

and if $X \sim P(\lambda)$ then

$$P[a \le X \le b] \cong \Phi\left(\frac{b+\frac{1}{2}-\lambda}{\sqrt{\lambda}}\right) - \Phi\left(\frac{a-\frac{1}{2}-\lambda}{\sqrt{\lambda}}\right)$$

21. Mean deviation about mean

$$MD\mu = E|X-\mu| = \sigma\sqrt{\frac{2}{\pi}} \cong 0.80\sigma$$

Example 2 : The average monthly sales of 5000 firms are normally distributed with mean Rs. 36 lakhs and S.D. Rs. 10 lakhs. Find

i) the number of firms the sales of which are below Rs. 30 lakhs.

ii) the percentage of firms the sales of which are more than Rs. 45 lakhs.

iii) the percentage of firms the sales of which are between Rs. 30 lakhs and Rs. 40 lakhs.

Solution : Let X be the random variable denoting the monthly sales of a firm (in lakhs rupees). Assuming that X follows the normal distribution with mean Rs. 36 lakhs and S.D. Rs. 10 lakhs we can compute.

(i) $P[X < 30] = P\left[Z < \frac{30-36}{10}\right] = P[Z < -0.6]$

$= \Phi(-0.6) = 1 - \Phi(0.6) = 1 - 0.7257 = 0.2743.$

Therefore, the number of firms the sales of which are below Rs. 30 lakhs = 5000 × 0.2743 = 1352.

(ii) $P[X > 45] = P[Z > \frac{45-36}{10}] = P[Z > 0.9]$

$= 1 - \Phi(0.9) = 1 - 0.8159 = 0.1841.$

So, the percentage of firms the sales of which are more than Rs. 45 lakhs. = 100 × 0.1841 = 18.41%.

(iii) $P[30 \le X \le 40] = \Phi\left(\frac{40-36}{10}\right) - \Phi\left(\frac{30-36}{10}\right)$

$= \Phi(0.4) - \Phi(-0.6) = \Phi(0.4) + \Phi(0.6) - 1.$

$= 0.6554 + 0.7257 - 1 = 0.3811.$

The percentage of firms the sales of which are between Rs. 30 laksh and Rs. 40 lakhs = 100 × 0.3811 = 38.11.

Example 3 : The average life of a certain type of small motor is 12 years with an s.d. of 2.5 years. The manufacturer replaces free all motors that fail while under guarantee. If he is willing to replace only 2.5% of the motors that fail, how long a guarantee should he offer? Assume that the lives of the motors follow a normal distribution.

Solution : Let X be the random variable denoting the life (in years) of a small motor. It is given that the life of a motor follows the normal distribution with mean $\mu = 12$ years and standard deviation, $\sigma = 2.5$ years. Let T be the guarantee period (in years) of a motor offered by the manufacturer so that he will replace 2.5% of motors.

i.e. P[life of a motor does not exceed its guarantee period] = 2.5/100 = 0.025.

or $P[X < T] = 0.025$

or $P[Z < (T - \mu)/\sigma] = 0.025$ $\phi(-1.960)$

or $T = \mu - 1.960\sigma = 12 - 2.5 \times 1.960 = 12 - 4.9$ 7.1 = 7 (app)

Therefore, the manufacturer should offer approximately 7 years of guarantee if he is willing to replace only 2.5% motors.

Example 4 : In a normal distribution 31% of the items are under 45 adn 8% are over 64. Find the mean and the standard deviation of the distribution.

Solution : Let X be a random variable which follows the normal distribution with mean μ and variance σ^2. It is given that

$P[X < 45] = 0.31$ and $P[X > 64] = 0.08$

Now, $P[X < 45\} = 0.31$

or, $P[(X - \mu)/\sigma < (45 - \mu)/\sigma] = 0.31$

or $\Phi[(45-\mu)/\sigma] = 0.31$

From the normal table $\Phi(-0.496) = 0.31$

So $(45 - \mu)/\sigma = -0.496$ ..... (1)

Again, $P[X > 64] = 0.08$

or, $P[X \leq 64] = 0.92$

or, $P[(X-\mu)/\sigma \leq (64 - \mu)/\sigma] = 0.92$

or, $\Phi[(64 - \mu)/\sigma] = 0.92$

From the normal table $\Phi(1.405) = 0.92$

So $(64 - \mu)/\sigma = 1.405$ ..... (2)

Solving (1) and (2), we get $\mu = 49.96 = 50$ (app.) and $\sigma = 10$.

Example 5 : A wholesale distributor of a product finds that the annual demand for the product is normally distributed with mean 120 and s.d. 16. If he orders only once a year, what quantity should be ordered to ensure that there is only a 5% chance of running short?

Solution : Let X be the random variable denoting the annual demand for the product. It is given that X follows the normal distribution with mean (μ) = 120 units and s.d. (σ) = 16 units. Suppose Q is the annual order quantity.

P[The annual demand exceeds the annual order quantity] = 0.05

or $P[X > Q] = 0.05$

or $P[X < Q] = 0.95$

or $P[(X - \mu)/\sigma \leq (Q - \mu)/\sigma] = 0.95$

or $\Phi[(Q - \mu)/\sigma] = 0.96$

From the normal table $\Phi(1.645) = 0.95$

So $(Q - \mu)/\sigma = 1.645$

or $Q = \mu + 1.645\sigma = 120 + 1.645 \times 16 = 146.32 = 146$ (app.)

Therefore, the yearly ordered quantity should be 146 (app.) to ensure that there is only a 5% chance of running short.

Example 6 : The mean purchases per day by a customer in a large store is Rs.250 with an s.d. of Rs. 100. If on a particular day, 100 customers purchased for Rs.378 or more, estimate the total number of customers who purchased from the store that day.

Solution : Let X be the random variable denoting the purchases (in rupees) per day by a certain customer in a large store. X follows the normal distribution with

mean = Rs. 250 and s.d. Rs. 100. Suppose N is the total number of customers who purchased from the store that day.

Given that,

P[a customer purchases more than Rs. 378 on a particular day] = 100/N

or, $P[X \geq 378] = 100/N$

or, $P[Z \geq (378 - 250)/100] = 100/N$

or, $P[Z < 128/100] = 1 - 100/N$

or, $\Phi(1.28) = 1 - 100/N = 0.9$ (from normal table)

So $N = 1000$.

Therefore, 1000 customers purchased from the store that day.

4.3 Exercises

1. Write down the probability density function of a normal distribution with mean μ and variance σ^2. Show that it is symmetric about μ.
2. State the important properties of the normal distribution.
3. Describe briefly the importance of the normal distribution in business decision.
4. Find the mode and points of inflexion of the normal distribution with mean μ and variance σ^2.
5. The income of a group of 1,00,00 persons was found to be normally distributed with mean Rs. 7500 p.m. and s.d. Rs. 500. What is the lowest income among the richest 1000?
6. A production engineer finds that on an average mechanic working in a machine shop completes a certain task in 30 minutes. The time required to complete the task is approximately normally distributed with an s.d. of 5 minutes. Find the probability that the task is completed a) in less than 15 minutes, b) in more than 20 minutes.
7. A bank manager finds that the lengths of times the customers have to wait for being attended to by the teller are normally distributed with mean 3 minutes

and s.d. of 0.6 min. Find the prob. that a customer has to wait

a) for less than 2 minutes.

b) for more than 1.5 minutes.

c) Between 1 and 2 minutes.

8. 1000 light bulbs with a mean life of 120 days are installed in a new factory, their length of life is normally distributed with an s.d. of 20 days.

a) How many bulbs will expire in less than 90 days?

b) If it is decided to replace all the bulbs together, what interval should be allowed between replacement, if not more than 10% should expire before replacement?

9. The mean of the inner diameters (in inches) of a sample of 200 tubes produced by a machine is 0.502 and the s.d. is 0.005. The purpose for which these tubes are intended allows a maximum tolerance in the diameter of 0.496 to 0.508 otherwise the tubes are considered defective. What percentage of the tubes produced by the machine is defective if the diameters are found to be normally distributed?

10. The Kolkata Municipal Corporation installed 2,000 bulbs in a street of Kolkata. If these bulbs have an average life of 1000 burning hours and standard deviation of 200 hours, what number of bulbs might be expected to fail between 700 and 1300 hours?

11. An editor of a publishing company calculates that it requires 10 months on an average to complete the publication process from manuscript to finished books with a standard deviation of 2.5 months. He believes that the distribution of publication time follows the normal law. Out of 350 books he will handle this year, how many will complete the process in less than a year?

12. In a normal distribution, 8% of the items are under 50 and 10% are over 60. Find the mean and standard deviation of the distribution.

13. What is a standard normal variate? Find out its mean and standard deviation.

14. The marks is Statistics are normaly distributed with mean 50 and S.D. 10. Find the proportion of individuals getting (a) 60 marks or more and (ii) less than 30 minutes.

It is given $\Phi(1) = 0.841745$.

Unit 5 ❑ Sampling Theory

Structure

5.1 Introduction

Sampling denotes the selection of a part of the aggregate with a view to obtain information about the whole. This aggregate or totality of statistical information on a particular character of all the members covered by an investigation is called population or universe. When the population size is very large, it may not be possible

to take a complete enumeration of the population. Then we select a small part of the population called sample and by examining this small part we can infer about the nature of the whole population. The basic objective of sampling is to make inference about the population by examining a small part of it. In other words, sampling is only a tool which helps us to know the characteristics of the population by examining only a small part of it.

Some application of sampling in business :

a) Sampling methods are used in market research for assessing customer behaviour, especially during launching of new products in the market.

b) Sampling is also used to estimate the proportion of defective incoming lots from suppliers.

c) In industry sampling is done for statistical quality control. During manufacture, a few consecutive items are picked from the production line at regular intervals of time and these items are thoroughly tested.

5.2 Some Basic Terms

Population : In statsitical application the term population is applied to any finite or infinite collection of individuals. It is practically synonymous with aggregate and does not necessarily refer to collation of living organisms. The population may be finite or infinite. By finite population we mean a population which contains a finite number of members. Similarly, by an infinite population we mean a population containing an infinite number of members. The population size is denoted by N.

Sample : A part of a population, or a sub-set from a set of units, which is provided by some process of selection with the object of investigating the proportion of the parent population. The sample size is denoted by n.

Census : The complete enumeration of a population or groups of a point of time with respect to well defined characteristic such as population, production, traffic on particular roads.

Sample survey : A survey which is carried out using a sampling method i.e., in which a portion of population is surveyed.

Sampling frame : A list, map or other specification of the units which constitute the available information relating to the population designated for a particular sampling scheme. That is, sampling frame is a complete list of elements comprising the population from which a sample is to be selected

Sampling distribution : The frequency distribution of all possible samples of a certain size drawn from a particular population.

Sampling error : This is the difference between population parameter and the observed statistic. The error which arises due to only a sample being used to estimate the population parameters is known as sampling error or sampling fluctuation.

This error is inherent and unavoidable in any and every sampling scheme. A sample with the smallest sampling error will always be considered a good representation of the population. This error can be reduced by increasing the sample size. When the sample survey becomes census, the sampling error becomes zero.

Non-sampling error : A sample estimate may be subject to other errors which, grouped together, are termed as non-sampling error. The main source the of non-sampling errors are:

i) Failure to measure some of the units in the selected sample.

ii) Observational error due to defective measurement techniques.

iii) Errors introduced in editing, coding and tabulating the results.

In practice, the census results may suffer from non-sampiing error although this may be from sampling error. The non-sampling error is likely to increase with increase in sample size, while sampling error decreases with the increase in the sample size.

Parameter: Any statistical measure computed based on all the units in the population is called a parameter, e.g., population mean (μ), s.d.(σ), proportion(p) etc.

Statistic : Any statistical measure computed on the basis of sample observations is called a statistic, e.g. sample mean ($\bar{x}$), s.d. (s), proportion (x/n) etc.

Estimator: An estimator is a rule or method of estimating a population parameter. It is generally expressed as a function of sample variates. An estimator is itself a random variable.

Estimate: A particular value of an estimator obtained from a set of values of a random sample is known as an estimate.

Random number table : A random number table is an arrangement of digits 0 to 9, in either a linear or rectangular pattern, where each position is filled with one of the digits. A table of random numbers is so constructed that all numbers 0, 1, 2, 3,..., 9 appear independent of each other. Some random number tables in common are :

i) Tippett's random number table

ii) Fisher and Yates' table.

iii) Kendall and Smith tables.

iv) A million random digits.

5.3 Procedure for Selecting a Random Sample

The simplest way of selecting a sample of the required size is by selecting a random number from 1 to N and then taking the unit bearing the number. The procedure involves a number of rejections since all numbers greater than N appearing in the table are not considered for selection. The use of a random number is modified, some of these modified procedures are :

i) Remainder approach

Let N be an r-digit number and let its r-digit highest multiple be N*. A random number R is chosen from 1 to N* and the unit with the serial number equal to the remainder obtained on dividing R by N is selected. If the remainder is zero, the last unit is selected. As an illustration suppose N = 32, the highest two-digit multiple of 32 is 96. For selecting a unit, one random number form 01 to 96 has to be selected. Let the random number selected be 75 the remainder is 11. Hence, the unit with serial number 11 is selected in the sample.

ii) Quotient approach

Let N be an r-digit number and let its r-digit highest multiple be N* such that N*/N=q . A random number R is chosen from 0 to (N* - 1). Dividing R by q the quotient r is obtained and the unit bearing the serial number (r - 1) is selected in the sample. As an illustration let N = 32 and hence

$N^* = 96$ and $q = 96/32 = 3$. Let the two-digit random number chosen be 65 which lies between 0 and 95. Dividing 65 by 3 the quotient is 21. Hence the unit bearing number (21 - 1) = 20 is selected in the sample.

5.4 Sampling Schemes

The procedure adopted to select the sample is known as sampling scheme.

Sampling schemes are broadly classified as non-probabilistic, probabilistic and mixed.

In non-probabilistic sampling, there is a fixed sampling rule but there is no probability attached to the mode of selection. On the other hand, if for each individual there is a definite pre-assigned probability of being selected, the sampling is said to be probabilistic. Probabilistic sampling is also called random sampling. In mixed sampling the selection process is partly probabilistic and partly non-probabilistic.

5.5 Some Random Sampling Schemes

5.5.1 Simple Random Sampling (SRS) :

The simplest most commonly used type of probability sampling is simple random sampling. In this sampling, each member of the population has the same probability of being included in the sample. Simple random sampling is said to be with or without replacement according as any individual once selected is returned to the population or not.

a) Simple Random Sampling with replacement (SRSWR) :

In this case, a unit is selected from a population with known probability and the unit is returned to the population before the next selection is made. Thus in this method at each selection, the population size remains constant and the probability at each selection or draw remains the same. Under this sampling scheme, a unit has chances of being selected more than once. Let us suppose that the size of the population is N and a sample of size n is to be drawn

from the given population. In case of SRSWR the n units of the sample are drawn from the population one by one. After each drawing the unit selected being returned to the population in such a way that at each drawing each of the N member gets the same probability 1/N of being selected, Clearly, here the same unit of the population may occur more than once in the sample and there will be N^n possible samples, each with probability $1/N^n$ to be selected.

b) Simple Random Sampling Without Replacement (SRSWOR) :

In this selection procedure, if a unit from a population of size N is selected, it is not returned to the population. Thus for any subsequent selection, the population size is reduced by one. Obviously, at the time of the first selection, the population size is N and the probability of a unit being selected randomly is 1/N, for the second unit to be randomly selected, the population size is (N - 1) and the probability of selection of any one of the remaining is 1/ (N - 1), similarly at the third draw, the probability of selection is 1/(N - 2) and so on. Here no member of the population can occur more than once in the sample. There are ${}^{N}C_{n}$ all possible samples each with probability

$$\frac{n}{N}\frac{n-1}{N-1}\frac{n-2}{N-2}\cdots\frac{1}{N-n+1}=\frac{1}{{}^{N}C_{n}}$$ to be selected.

5.5.2 Stratified Random Sampling

In stratified sampling the population of N units is sub-divided into k sub-population called strata, the i-th sub-population having Nj units (j = 1, 2, ..., k). These sub-populations are non-overlapping so that they comprise the whole population such that $N_1 + N_2 + \ldots + N_k = N$.

A sample is drawn from each stratum independently, the sample size within the i-th stratum being n_i (i = 1, 2, ... , k) such that

$$n_1 + n_2 + \ldots + n_k = n.$$

The procedure of taking samples in this way is known as stratified sampling. If the sample is taken randomly from each stratum, the procedure is known as stratified random sampling. The stratification of population should be done in such a way that the strata are homogeneous within themselves, with respect to the characteristic under study.

For example, human population may be divided into different strata or sub-groups on the basis of sex, age group, education, income and occupation etc.

5.5.3 Multistage Sampling

This sampling involves the selection of units in more than one stage. In such a sampling, the population consists of a number of first stage units or primary sampling units (psu), then a sample is taken of the second stage units. This process continues until the selection of the final sampling units.

Suppose a sample of 10,000 urban households from all over the country is to be selected. In such a case the first stage sample may involve the selection of states. The second stage may involve the selection of districts. The third stage may involve selection of cities. Some wards from each selected city may be chosen. Finally, some households from each ward may be selected. Thus 10,000 urban households are arrived at in five stages. So the final stage unit is the household. Thus this sampling becomes five stage sampling :

1st stage units - states

2nd stage units - districts

3rd stage units - cities

4th stage units - wards

5th stage and final stage units - households.

5.5.4 Cluster Sampling or Area Sampling

In random sampling, it is presumed that the population has been divided into a finite number of distinct and identifiable units defined as sampling units. The smallest unit into which the population can be divided is called an element of the population. A group of such elements is known as a cluster. When the sampling unit is a cluster, the procedure is called cluster sampling. If the entire area containing the population under study is divided into smaller segments and each element in the population belongs to one and only one segment, the procedure is some times called area sampling. As a simple rule, the number of elements in a cluster should be small and the number of cluster should be large. After dividing the population into specified cluster, the required number of clusters can be selected either by equal or unequal probabilities of selection. All the elements in selected cluster are enumerated.

The advantages of cluster sampling are as follows :

a) Collection of data for neighbouring elements is easier, cheaper, faster and operationally more convenient than observing units spread over a region.

b) It is less costly than simple random sampling.

c) When the sampling frame of elements may not be readily available, ideally clusters should be formed so that within a given cluster study objects are as heterogeneous as possible. Each cluster should have a complete representation of the study objects and identical to every other cluster.

5.5.5 Systematic Sampling

It is different from the SRS mainly on the basis of the fact that every combination of study objects does not possess equal chance of being selected. In particular, suppose that a population has N objects and a sample of size n is to be selected. Now $k = N/n$ is called the sampling interval. Now a random number between 1 to k is drawn. Thus in systematic sampling the study objects to be included are r, r + k, r + 2k, ..., r + (n - 1)k. If N = nk, the resultant sample is called every k-th systematic and such a procedure is known as **linear systematic sampling.**

If $N \neq nk$ and every k-th unit be included in a circular manner till the whole list is exhausted, will be called **circular systematic sampling.**

To overcome the difficulty of varying sample size under the situation $N \neq nk$, the procedure is modified slightly by which a sample of constant size is always obtained. The procedure consists in selecting a unit, by a random start, from 1 to N and then selecting every k-th unit, k being an integer nearest to N/n, in a circular manner, until a sample of n units is obtained. Suppose that a unit with random number i is selected. The sample will then consist of the units corresponding to the serial numbers

$i + jk$, if $i + jk \leq N$ for $j = 0, 1, 2 \ldots n - 1$

$i + jk - N$, if $i + jk > N$

As an illustration, let N = 11 and n = 4. Then k = 3.

The possible samples are:

(1, 4, 7, 10), (2, 5, 8, 11), (3, 6, 9, 1), (4, 7, 10, 2), (5, 8, 11, 3), (6, 9, 1, 4),

(7, 10, 2, 5), (8, 11, 3, 6), (9, 1, 4, 7), (10, 2, 5, 8) and (11, 3, 6, 9).

Advantages of Systematic Sampling :

The main advantage of the systematic sampling is its simplicity of selection, operational convenience and every spread of sample over the population. It has, therefore, been very useful in forest survey for estimating the volume of timber, in fisheries for estimating total catch of fish etc. Another advantage is that, except for population with periodicities, systematic sampling provides an effective estimate as compared to alternative design. Some times, systematic sampling variances are much smaller than the variance for random selection of units within strata.

5.6 Non-Probability Sampling Schemes

We have so far discussed probabilistic sampling schemes. In reality, because of various difficulties involved in obtaining reliable lists of the desired target population, it is difficult to use a probabilistic scheme. Therefore, some compromises could be made, some of the non-probabilistic techniques may also be made explicitly in cases where it is not feasible to use probability-based methods. The major difference is that in non-probabilistic techniques the extent of bias in selecting sample is not known. This makes it difficult to say anything about the representativeness or accuracy of the sample. There are four major non-probabilistic sampling schemes. They are convenience sampling, judgmental or purposive sampling, quota sampling and snowball sampling.

5.6.1 Snowball Sampling

This technique is used where the population being sought is a small one and chances of finding them by traditional methods are low. For example, one respondent being used to generate names of others is called snowballing and it can be again done on the second set of respondents. It could be also called networking to find respondents.

For example, to find owner of Mercedes benz car in a city, we may go to one or two and ask them if they know any one else who owns the car.

5.6.2 Convenience Sampling

It refers to selecting a sample of study objects based on convenience. Thus a study may include objects which are conveniently located, willing to co-operate in offering the necessary data and in the process one would derive the advantage of economy in cost or time. Findings based on convenient sampling procedure can not be generated. In exploratory type of results, convenience sampling procedures may be adopted in conducting the focus group interview or survey. Similarly, questionnaire be pretested on a sample selected by convenience. Such sampling scheme helps to understand the possible variability of responses within a short span of time and cost. An example of convenience sampling includes on the T.V. reporters who catch any person passing and interviewed on the street.

5.6.3 Purposive or Judgmental Sampling

When the choice of individual items of a sample entirely depends on the individual judgment of the investigator, it is called purposive or judgmental sampling.

In this method, the units constituting the sample are chosen not according to some scientific procedure, but according to personal choice of the person who selects the sample. Two or more such independent purposive samples may give widely different estimates of the same population. For example, an observer who wishes to take a sample of oranges from a lot run his eyes over the whole lot and then chooses average oranges, averages in size, shape, or whatever other quality he may have in his mind.

5.6.4 Quota Sampling

If the sampling frames for the different strata into which the population may be divided are not available and are costly to construct, it may be possible to fix up a sample quota for each stratum and to continue sampling,until the necessary quota for each strata is filled up. The objective is to gain the benefits of stratifications as far as possible without the high costs that may be incurred in any other to have recourse to probabilistic sampling. The method has been found useful in many socio-economic and opinion survey. In many studies the researcher can, a-priori, decide on the type of target respondent and quotas of different groups of respondents.

Suppose in a certain region we want to conduct a survey of households where total number of households is 1,00,000. It is required that a sample of 1 percent, i.e. 1,000 households is to be covered. A sample of 1,000 households has been chosen, subject to the condition that 600 of those should be from rural areas and 400 from the urban areas. Likewise, of the 1,000 households the rich households should number 75, the middle class ones 325 and the remaining 600 should be from poor class.

Advantages :

a) It is economical as travelling costs can be reduced.

b) It is administratively convenient.

c) When the field work is to be done quickly, quota sampling is the most appropriate and feasible.

d) It is independent of existence of sample frames. Whenever a suitable sampling frame is not available, quota sampling is perhaps the only choice available.

Limitations :

a) Since the quota sampling is not a random selection, it is not possible to calculate the estimates of standard error for the sample results.

b) This is not, in general, a representative sample, it depends entirely on the mood and convenience of the interviewer.

c) It may be extremely difficult to supervise and control the field survey under quota sampling.

5.7 Exercises

1. What are the advantages of sample surveys over complete census?
2. What are random sampling numbers ?
3. Describe the following methods of sampling with suitable business examples, i) Stratified sampling ii) Multistage sampling, iii) Systematic sampling and (iv) Purposive sampling.
4. What are the differences between sampling error and non-sampling error? How car you control non-sampling error?

5. In what respect does circular systematic sampling differ from linear systematic sampling? When are these two types of sampling equivalent?
6. Distinguish between sampling with replacement and sampling without replacement.
7. What is the standard error of a statistic and what is its uility? What is the difference between standard error and standard deviation?
8. Distinguish between a parameter and a statistic. Which one of these is a variable and why?
9. Describe various methods of drawing a random sample from a finite population.
10. Draw a random sample of size 10 without replacement from the following data on daily sales (in thousand rupees) of 32 shops in Kolkata, stating clearly the procedure followed by you.

35	28	27	33	47	28	40	35
24	32	50	26	38	36	37	41
26	35	46	41	43	33	46	26
45	46	48	27	36	41	32	30

You may use the random number given below :

5967	8941	7889	3335	7577	9735
3042	8409	7053	5364	5872	1143

11. What is meant by stratified random sampling? Explain the procedure and advantages of stratification.
12. A population contains six units given below : 2, 6, 5, 1, 7 and 3. Consider all possible samples of size two and verify that the mean of the population is exactly equal to the mean of the sample means.
13. A simple random simple of size 64 is drawn from a finite population consisting of 122 units. If the population s.d. is 16.8, find the standard error of the sample mean when the sample is drown (i) with replacement and (ii) without replacement.
14. A simple random sample of size 10 is drawn without replacement from a finite population consisting of 200 units. If the number of defective units in the population is 15, find the S.E. of the proportion of defectives.

Unit 6 ❑ Sampling Distribution

Structure

6.1 Introduction

The statistical measures calculated on the basis of the population observations are called parameters and the statistical measures calculated on the basis of the sample observations are called statistic. For a given population a parameter has always a fixed value. But since different samples can be drawn from the same population, the value of the statistic is likely to vary from one sample to another. The differences of statistic are called sampling fluctuations. Thus, if a number of samples, each of size 'n' are taken from the same population and if for each sample the value of a statistic

is calculated, a series of values of the statistic will be obtained. If the number of samples be large, these may be arranged in a frequency distribution table. The frequency distribution of the statistic is called the sampling distribution of the statistic. Therefore, sampling distribution of a statistic may be defined as the probability law which the statistic follows, if repeated random samples of a fixed size are drawn from a specified population. Standard error of a statistic is the standard deviation calculated from the sampling distribution of the statistic.

Unbiased Estimator : An estimator T is said to be an unbiased estimator of a parameter θ if $E(T) = 0. \forall \theta$

If $E(T) \neq 0$, then T is said to be a biased estimator of θ. The bias of the estimator is given by

Bias = $E(T) - \theta$ or $E(T) + \theta$.

Let us consider a random sample of size n as $x_1, x_2, ..., x_n$ drawn from a population with mean μ and variance σ^2. Then we have,

$E((\bar{X})) = \mu$ i.e. the sample mean $\bar{x}$, is an unbiased estimator of the population mean μ.

$E(S^2) = \frac{n-1}{n}\sigma^2 \neq \sigma^2$ i.e. the sample variance S^2 is not an unbiased estimator of population varance, σ^2.

6.2 Sampling Distribution of the Sample Mean

Let us consider a random sample of size n as $x_1, x_2, ... x_n$ drawn from a normal population with mean μ and variance σ^2. If $\bar{x}$ denotes the sample mean which is defined by $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ follows a normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$ That is , $Z = \frac{(\bar{x}-\mu)}{\sigma/\sqrt{n}}$ follows the standard normal distribution.

6.3 Chi-square (χ^2 Distribution)

A random variable y is said to follow the Chi-square (χ^2) distribution with n degrees of freeom if its p.d.f. is of the form.

$f(y) = \text{Constant } e^{-y/2} y^{(n/2)-1}; \ 0 < y < \infty.$

Again, if X is a random variable which follows the normal distribution with mean μ and standard deviation σ, then $Z = \frac{x-\mu}{\sigma}$ is a standard normal variate. The square of Z i.e. $Z^2 = \frac{(x-\mu)^2}{\sigma^2}$ follows a χ^2 square variate with one degree of freedom and is written as χ_1^2 and the range of χ^2 distribution is from 0 to ∞.

Result 6.1 : If $Z_1, Z_2, \ldots\ldots, Z_n$ are n independent standard normal variables, then

$\sum_{i=1}^{n} Z_i^2 \sim \chi_n^2, 0 < \chi^2 < \infty.$

Result 6.2 : If X is a random variable which follows the normal distribution with mean μ and standard deviation σ, then

$\sum_{i=1}^{n} (x_i - \mu)^2 / \sigma^2$ follows the chi-square distribution with n degrees of freedom ; and

$\sum_{i=1}^{n} (x_i - \bar{x})^2 / \sigma^2$ follows the chi-square distribution with (n–1) degrees of freedom.

6.3.1 Properties of χ^2 distribution

1. The distribution has only one parameter, i.e., number of degrees of freedom which is a positive integer.
2. χ^2 distribution realizes only non-negatie values.
3. The mean of the distribution = $E(\chi_n^2) = n$ and the variance of the distribution $= V(\chi_n^2) = 2n$.
4. χ^2 distribution is not symmetric. It is positively skewed.

5. If χ_1^2 and χ_2^2 are two chi-square variates with n_1 and n_2 degrees of freedom respectively, then $\chi_1^2 + \chi_2^2$ follows also a chi-square distribution with $(n_1 + n_2)$ degrees of freedom.

6.4 The Student's t-Distribution

A random variable is said to follow Student's t distribution or simply t distribution with n degrees of freedom, if its p.d.f. is of the form

$$f(t) = \text{Constant.} \left(1+\frac{t^2}{n}\right)^{-(n+1)/2}, -\infty < t < \infty$$

Let $x_1, x_2, \ldots\ldots, x_n$ be n independent random variables from a normal population with mean μ and standard deviation σ (unknown).

When σ is not known, it is estimated by the sample standard deviation $s = \sqrt{\frac{1}{n-1}\sum(x_i-\bar{x})^2}$ which is an unbiased estimator of σ. In such a case we would like to know the exact distribution of the statistic $\frac{\sqrt{n}(x-\mu)}{s}$ and the answer to this is provided by the t-distribution.

W.S. Gosset defined the t statistic as $t = \frac{\sqrt{n}(\bar{x}-\mu)}{s}$ which follows the t-distribution with (n–1) degrees of freedom.

Result 6.3 : If Z and Y are independent random variables, where Z follows the standard normal distribution and Y follows the chi-square distribution with n degrees of freedom, then $t = \frac{Z}{\sqrt{Y/n}}$ follows the t distribution with n degrees of freedom.

Result 6.4 : If $x_1, x_2, \ldots\ldots, x_n$ be n independent random variables from a normal population with mean μ and standard deviation σ (unknown), then $t = \frac{\sqrt{n}(\bar{x}-\mu)}{s}$ which follows the t-distribution with (n–1) degrees of freedom.

Proof : To show this we can write $t = \frac{\bar{x}-\mu}{s/\sqrt{n}} = \frac{\sqrt{n}(\bar{x}-\mu)}{s}$

Dividing the numerator and the denominator by σ, we get

$$t = \frac{\frac{\sqrt{n}(\bar{x}-\mu)}{\sigma}}{\frac{s}{\sigma}} = \frac{\frac{(\bar{x}-\mu)}{\sigma/\sqrt{n}}}{\sqrt{s^2/\sigma^2}} = \frac{\frac{(\bar{x}-\mu)}{\sigma/\sqrt{n}}}{\sqrt{\frac{1}{n-1}\frac{\sum(x_i-\bar{x})^2}{\sigma^2}}}$$

$$= \frac{\frac{(\bar{x}-\mu)}{\sigma/\sqrt{n}}}{\sqrt{\frac{\chi^2_{n-1}}{n-1}}} = \frac{\text{Standard Normal Variate}}{\sqrt{\frac{\chi^2\text{ - Variate with } (n-1) \text{ degrees of freedom}}{(n-1)}}}$$

Result 6.5 : If two independent random samples of sizes n_1 and n_2 from two normal populations with means μ_1 and μ_2 and common variance σ^2, are taken then

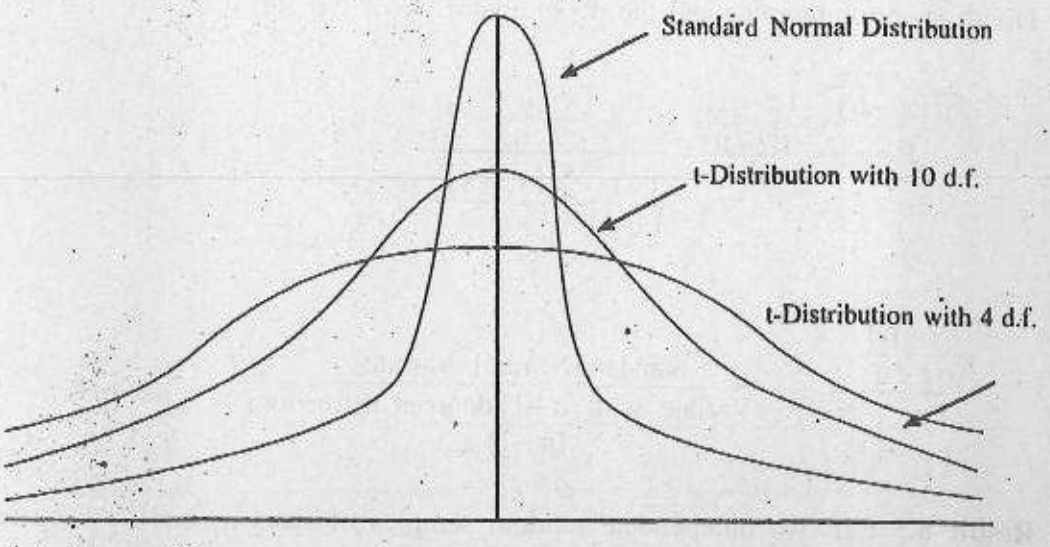
$$t = \frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{s\sqrt{1/n_1+1/n_2}}$$

which follows the t-distribution with $(n_1 + n_2 - 2)$ degrees of freedom, where $s^2 = \frac{n_1S_1^2+n_1S_2^2}{n_1+n_2-2}$ with $S_1^2 = \frac{1}{n_1}\sum_{i=1}^{n_1}(n_{1i}-\bar{x}_1)^2$ and $S_2^2 = \frac{1}{n_2}\sum_{i=1}^{n_2}(n_{2i}-\bar{x}_2)^2$ be the variances of the first sample and the second sample respectively.

6.4.1 Properties of the t-Distribution

1. Like χ^2 - distribution the t-distribution also has one parameter $v = (n-1)$ where n denotes sample size. Hence, this distribution is known if n is known.
2. Mean of the random variable t is zero and standard deviation is $\sqrt{\frac{v}{v-2}}$, for $v > 2$.
3. The probability curve of the t-distribution is symmetrical about the ordinate at $t = 0$. Like a normal variable, the t variable can take any value from $-\infty$ to ∞.

4. The disribution approaches the normal distribution as the number of degrees of freedom becomes large.



5. The random variable t is defined as the ratio of a standard normal variate to the square root of χ^2 – variate divided by its degrees of freedom.

6.5 Snedecor's F-Distribution

A random variable is said to follow the F distribution with n_1 and n_2 degrees of freedom if the p.d.f. is of the form

$$f(F) = \text{Constant } F^{(n_1/2)-1}\left(1+\frac{n_1}{n_1}F\right)^{(n_1+n_1)/2}, 0 < F < \infty.$$

Result 6.6 : If Y_1 and Y_2 are independent random variables, where Y_1 and Y_2 follow chi-square distributions with n_1 and n_2 degrees of freedom respectively, then $F = \frac{Y_1/n_1}{Y_2/n_2}$ follows the F distribution with n_1 and n_2 degrees of freedom.

Result 6.7 : Let there be two independent random samples of sizes n_1 and n_2 from two normal populations with variances σ_1^2 and σ_2^2 respectively. Further, let

$s_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)^2$ and $s_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(x_{2i} - \bar{x}_2)^2$ be the variances of the first sample and the second sample respectively. Then the F-statistic is defined as the ratio of two χ^2 – variates. Thus, we can write

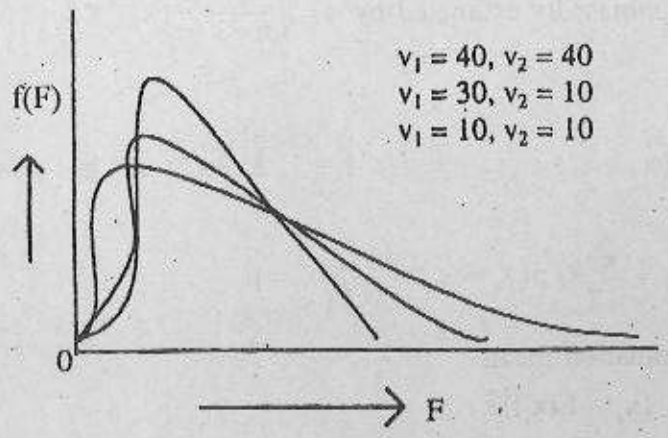
$$F = \frac{\chi^2_{(n_1-1)}/(n_1-1)}{\chi^2_{(n_2-1)}/(n_2-1)} = \frac{\dfrac{(n_1-1)s_1^2}{\sigma_1^2}\Big/(n_1-1)}{\dfrac{(n_2-1)s_2^2}{\sigma_2^2}\Big/(n_2-1)} = \frac{\dfrac{s_1^2}{\sigma_1^2}}{\dfrac{s_2^2}{\sigma_2^2}} \sim F_{n_1-1,\, n_2-1}$$

6.5.1 Properties of the F-Distribution

1. This distribution has two parameters v_1 ($= n_1 - 1$) and v_2 ($= n_2 - 1$).
2. The mean of the F – variate with v_1 and v_2 degrees of freedom is $v_2/(v_2-2)$

   and Standard Deviation $= \left(\frac{v_2}{v_2-1}\right)\sqrt{\frac{2(v_1+v_2-2)}{v_1(v_2-4)}}$

   We note that the mean will exist if $v_2 > 2$ and the standard error will exist if $v_2 > 4$. Further, the mean > 1.
3. The random variate F can take only values from 0 to ∞. The curve is positively skewed.
4. For large values of v_1 and v_2 the distribution approaches the normal distribution. This behaviour is shown in the following figure.



6.6 Sampling distribution of the Sample Mean for Non-normal Population

Let there be a population containing N units. Let us consider a random sample x_1, x_2, ..., x_n of size n drawn from this population. Then the sample mean is defined by $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Now we are interested in the sampling distribution of the statistic $\bar{x}$. When sampling is done with replacement, we get n^n equally likely possible samples and when sampling is done without replacement then the possible number of samples will be ${}^{N}C_{n}$. Here samples will be all distinct but sample means may not be all distinct. The probability distribution of the sample mean $\bar{X}$ with corresponding probability is called sampling distribution of the sample mean.

Case – 1 : SRSWR

Result 6.8 : It x_1, x_2, ... x_n be a simple random sample of size n drawn with replacement from a finite population of size N with $E(x_i) = \mu$ and $V(x_i) = \sigma^2$, then

(i) $E(\bar{x}) = E\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \mu$ and

(ii) $V(\bar{x}) = V\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \sigma^2/n$

Note : σ^2 is unbiasedly estimated by $s = \sqrt{\frac{1}{n-1}\sum(x_i - \bar{x})^2}$.

Proof :

Here $P(x_y = x_i) = \frac{1}{N}$, $\quad i = 1, 2, 3, ..., N ; r = 1, 2, 3, ..., n.$

So $E(x_r) = \sum_{i=1}^{N} x_i \, p(x_r = x_i) = \frac{1}{N}\sum_{i=1}^{N} x_i = \mu$

where μ = population mean

and $V(x_i) = E\{x_i - E(x_i)\}^2$

$$= E(x_r - \mu)^2$$

$$= \sum_{i=1}^{N}(X_i-\mu)^2 P(\bar{x}=x_i)$$

$$= \frac{1}{N}\sum_{i=1}^{N}(X_i-\mu)^2$$

$$= \sigma^2$$

where σ^2 is the population variance.

Now (i) $E(\bar{x}) = E\left(\frac{1}{n}\sum_{i=1}^{n}x_i\right)$

$$= \frac{1}{n}\sum_{i=1}^{n}E(x_i)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\mu$$

$$= \frac{1}{n}n\mu$$

$$= \mu$$

(ii) $V(\bar{x}) = V\left(\frac{1}{n}\sum_{i=1}^{n}x_i\right)$

$$= \frac{1}{n^2}\sum_{i=1}^{n}V(x_i) \text{ (Since } x_1\text{ 's are independent)}$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2$$

$$= \frac{1}{n^2}n\sigma^2$$

$$= \frac{\sigma^2}{n}$$

Thus in SRSWR the standard error of the sample mean is

$$SE(\bar{x}) = \frac{\sigma}{\sqrt{n}}$$

SASE - II : SRSWOR

Result 6.9 :

Let $x_1, x_2, \ldots\ldots, x_n$ be a simple random sample of size n drawn without replacement from a finite population of size N with $E(x_i) = m$ and $V(x_i) = \sigma^2$. Then

(i) $E(\bar{x}) = E\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \mu$ and

(ii) $V(\bar{x}) = V\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{N-n}{N-1}\frac{\sigma^2}{n}$.

Note : $\frac{N-n}{N-1}\frac{\sigma^2}{n}$ is unbiasedly estimated by $\frac{N-n}{N-1}\frac{s^2}{n}$

Proof :

Here $P(x_r = x_i) = \frac{1}{N}$

and $P(x_r = X_i,\ x_s = X_j) = \frac{{}^{N-2}P_{n-2}}{{}^{N}P_{n}} = \frac{1}{N(N-1)}$

Now, COV. $(x_r\ x_s) = \sum\sum_{i\neq j}(X_i-\mu)(X_j-\mu)P(x_r = x_i,\ x_s = x_j)$

$= \frac{1}{N(N-1)}\sum\sum_{i\neq j}(X_i-\mu)(X_j-\mu)$

$= \frac{1}{N(N-1)}\left[\left\{\sum_{i=1}^{N}(X_i-\mu)\right\} - \sum_{i=1}^{N}(X_i-\mu)^2\right]$

$= \frac{1}{N(N-1)}[0 - n\sigma^2] = -\frac{\sigma^2}{N-1}$

$E(\bar{x}) = \mu$ same as SRSWR

Also, $V(\bar{x}) = V\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)$

$$= \frac{1}{n^2} V\left(\sum_{r=1}^{n} x_r\right)$$

$$= \frac{1}{n^2}\left[\sum_{r=1}^{n} V(x_r) + \sum\sum_{r \neq s} Cov(x_r, x_s)\right]$$

$$= \frac{1}{n^2}\left[\sum_{r=1}^{n} \sigma^2 + \sum\sum\left(-\frac{\sigma^2}{N-1}\right)\right]$$

$$= \frac{1}{n^2}\left[n\sigma^2 - \frac{n(n-1)}{N-1}\sigma^2\right]$$

$$= \frac{\sigma^2}{n}\,\frac{N-n}{N-1}$$

Thus in SRSWOR, the standard error of the sample mean is

$$SE(\bar{x}) = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

The term $\sqrt{\frac{N-n}{N-1}}$ is termed as finite population correction (f.p.c.)

Example 1 : Construct a sampling distribution of the sample mean for the following population when random samples of size 2 are taken from it (a) with replacement and (b) without replacement. Also find the mean and standard error of the distribution in each case.

Poplulation Unit :	1	2	3	4
Observation :	22	24	26	28

Solution :

The mean and standard deviation of the population are

$$\mu = \frac{22+24+26+28}{4} = 25 \text{ and}$$

$$\sigma = \sqrt{\frac{(22)^2+(24)^2+(26)^2+(28)^2}{4} - (25)^2} = \sqrt{5} = 2.236 \text{ respectively.}$$

(a) When random samples of size 2 are drawn, we have $4^2 = 16$ samples shown below :

Sample No.	Sample Values	$\bar{X}$
1	22, 22	22
2	22, 24	23
3	22, 26	24
4	22, 28	25
5	24, 22	23
6	24, 24	24
7	24, 26	25
8	24, 28	26
9	26,22	24
10	26, 24	25
11	26, 26	26
12	26, 28	27
13	28, 22	25
14	28, 24	26
15	28, 26	27
16	28, 28	28

Since all of the above samples are equally likely, therefore, the probability each value of $\bar{x}$ is $\frac{1}{16}$. Thus, we can write the sampling distribution of $\bar{x}$ as given below :

$\bar{x}$	22	23	24	25	26	27	28	Total
$p(\bar{x})$	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{3}{16}$	$\frac{4}{16}$	$\frac{3}{16}$	$\frac{2}{16}$	$\frac{1}{16}$	1

The mean of $\bar{X}$, i.e.,

$$\mu_{\bar{x}} = E(\bar{x}) = 22\times\frac{1}{16} + 23\times\frac{2}{16} + 24\times\frac{3}{16} + 25\times\frac{4}{16} + 26\times\frac{3}{16} + 27\times\frac{2}{16} + 28\times\frac{1}{16} = 25$$

Further, S.E. of $(\bar{x}) = \sigma_{\bar{x}} = \sqrt{E(\bar{x})^2 - [E(\bar{x})]^2}$, where

$$E(\bar{x}^2) = \frac{1}{16}(22^2 + 23^2 \times 2 + 24^2 \times 3 + 25^2 \times 4 + 26^2 \times 3 + 27^2 \times 2 + 28)$$

$$= 627.5$$

Thus, $\sigma_{\bar{x}} = \sqrt{627.5 - 25^2} = \sqrt{2.5}$ which is equal to $\frac{\sigma}{\sqrt{n}}$.

(b) When random samples of size 2 are drawn without replacement, we have ${}^4C_2$ samples shown below :

Sample No.	Sample Values	$\bar{X}$
1	22,24	23
2	22, 26	24
3	22, 28	25
4	24, 26	25
5	24, 28	26
6	24, 28	27

Since all the samples are equally likely, the probability of each value of $\bar{x}$ is $\frac{1}{16}$. Thus, we can write the sampling distribution of $\bar{X}$ as

$\bar{x}$	23	24	25	26	27	Total
$p(\bar{x})$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{2}{6}$	$\frac{2}{6}$	1

Further, $\mu = E(\bar{x}) = \frac{1}{6}[23 + 24 + 25 \times 2 + 26 + 27] = 25$

To find S.E. $E(\bar{x})$, we first find $E(\bar{x}^2)$

$$E(\bar{x}^2) = \frac{1}{6}(23^2 + 24^2 + 2 \times 25^2 + 26^2 + 27) = \frac{3760}{6} = 626.67$$

Thus, $\sigma_{\bar{x}} = \sqrt{626.67 - 25^2} = \sqrt{1.67} = 1.292$

Alternatively, $\sigma_{\bar{x}} = \sqrt{\frac{N-n}{N-1}\frac{\sigma^2}{n}} = \sqrt{\frac{4-2}{3} \times \frac{5}{2}} = \sqrt{1.67} = 1.29$

6.7 The Central Limit Theorem

According to the central limit theorem, for a large sample size, the sampling distribution of the sample mean $\bar{x}$ is approximately normal, regardless of the shape of the population distribution. Symbolically, the mean of the sampling distribution of $\bar{x}$ is μ and the standard deviation is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

By a large sample size, we mean that $n \geq 30$.

6.8 Sampling Distribution of Sample Proportion

Let the parameter p denote the proportion of successes in a binomial population. Further, let p denote the sample proportion of successes in n trials. We know from the central fimit theorem that the sampling distribution of p will be approximately normal with mean p and standard error. In other words,

let π = population proportion of units belonging to a particular class

and p = sample proportion of units belonging a particular class.

Result 6.10 : SRSWR

If π = proportion of units belonging to a particular class in an infinite population. Then,

(i) $E(p) = \pi$ and

(ii) $V(p) = \frac{\pi(1-\pi)}{n}$

Note : $\frac{\pi(1-\pi)}{n}$ is unbiasedly estimated by $\frac{p(1-p)}{n}$.

Result 6.11 : SRSWOR

If π = proportion of units belonging to a particular class in a finite population of

size N. Then

(i) $E(p) = \pi$ and

(ii) $V(p) = \frac{N-n}{N-1}\frac{\pi(1-\pi)}{n}$

Note : $\frac{N-n}{N-1}\frac{\pi(1-\pi)}{n}$ is unbiasedly estimated by $\frac{N-n}{N}\frac{p(1-p)}{n-1}$

6.9 Sampling Distribution of the Difference Between Two Sample Means

If two independent random samples of sizes n_1 and n_2 from two normal populations with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2, are taken then

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n^1} + \frac{\sigma_1^2}{n_1}}}$$

follows the standard normal distribution.

6.10 Sampling Distribution of the Difference Between Two Sample Proportions

Let π_i = proportion of units belonging to a particular calss in the i-th population, i = 1, 2, and p_i = sample proportion of units belonging to a particular class in a sample of size n_i.

If two populations are independent, then for large sample sizes

$$Z = \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}}$$

follows the standard normal distribnution.

6.11 Exercises

1. If x_1, x_2 and x_3 be a random sample from $N(0, \sigma^2)$ population, what is the distribution of $(x_1^2 + x_2^2 + x_3^2) / \sigma_2$? State the sampling distribution of the statistic

   $\sqrt{2x_1} / \sqrt{(x_2^2 + x_2^2)}$ and $x_1^2 + x_2^2$.

2. Derive the formula for the expectation and standard error of the sample proportion in both SRSWR or SRSWOR. Show that the standard error can not exceed the values $1/2\sqrt{n}$ and $(1/2\sqrt{n})\ \sqrt{(N-n)/(N-1)}$.

3. A random sample of size 17 from an $N(\mu, \sigma^2)$ yielded a sample variance of 25. What is the probability that the sample mean will not differ from the population mean by 21:18 in absolute value?

4. If two independent random samples of sizes 10 and 12 are taken from $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ respectively, where $\sigma_1^2 = 4$ and $\sigma_2^2 = 9$, what is the probability that $\bar{X} - \bar{Y}$ will not differ from $\mu_1 - \mu_2$ by 3 in absolute value?

5. Obtain the expectation and standard error of sample mean for a random sample of size n drawn from a population of size N by (a) SRSWR and (b) SRSWOR.

6. Starting from the density function of a normal distribution, state density function of χ^2, t and F-distributions. Also state their important properties.

7. A population consists of numbers 4, 7 and 9.

   (a) Enumerate all possible samples of size 2 which can be drawn from the population without replacement.

   (b) Show that the mean of the sampling distribution of the sample means is equal to the population mean.

c) Calculate the variance of the sample mean and show that it is less than the population variance.

8. The income of a group of workers is distributed normally with mean Rs.3000 and standard deviation of Rs. 500. If a random sample of 49 workers is taken from this group, determine

   a) the sampling distribution of mean,

   b) the probability that the sample mean will lie between Rs. 2800 and Rs. 3300 and

   c) the number of samples having their means greater than Rs. 3200.

9. The life of tyres manufactured by a company A is distributed normally with mean 16,000 kms and s.d. 2,000 kms and of that manufactured by a company B is distributed normally with mean 20,000 kms and s.d. 2500 kms. If 64 tyres of company A and 36 tyres of company B are selected at random, determine the sampling distribution of the difference between mean life of tyres.

10. 10% of articles produced by Machine A are defective and 5% of them produced by Machine B are defective. A random sample of 250 articles is taken from Machine A's output and a random sample of 300 articles is taken from Machine B's output. What is the probability that the difference in sample proportion of defective articles is greater than or equal to 0.03?

11. An automatic machine pours, on an average 199 ml. of soft drink into bottles that are supposed to contain 200 ml. The standard deviation of the amount filled is 3 ml. Assuming that the amount filled is distributed normally, find the probability that a random sample of 25 bottles will have a mean of 200 ml.or more.

12. Distinguish between

    a) Parameter and statistic.

    b) Sampling distribution and probability distribution.

    c) Standard deviation and standard error.

    Explain the concept of sampling distribution of a statistic.

13. If x_1, x_2 and x_3 is a simple random sample of size 3 from a large population with mean 12 and variance 9, evaluate the expected value and standard error of the statistic $T = (2x_1 + x_2 - 3x_3)$.

14. A simple random sample of size 36 is drawn from a finite population consisting of 101 units. If the population S.D. is 12.6, find the standard error of sample

mean when the sample is drawn (i) with replacement and (ii) without replacement.

15. The diameter of a compound produced on a semi-automatic machine is known to be distributed normally with mean of 10 mm and standard deviation of 0.2 mm. If we pick up a random sample of size 16, what is the probability that the sample mean will lie between 9.95 to 10.05 mm?

16. It is known that 7% of the bolts manufactured by a factory are defective. If a random sample of 100 bolts are chosen at random from a day's production, obtain the sampling distribution of (i) the number of defective and ii) the proportion of defective bolts.

17. The guaranteed life of a certain type of electric bulbs is 1000 hours with standard deviation of 125 hours. It is proposed to sample the output so as to assure that 90% of the bulbs do not fall short of the guaranteed average life by more than 2.5%. What should be the minimum size of the sample?

    [Hint: $P(\bar{X} - \mu > -2.5\% \text{ of } \mu) = 0.90$. From the area of the standard normal distribution $\sqrt{n} = 5 \times 1.28$ or $n = 40.92 = 41$ (app.)]

18. A lot of 100 items contains 20 defectives. If a simple random sample of size 10 is drawn without replacement, find out the standard error of the sample proportion of defective items.

19. What is meant by stratified random sampling? Explain the procedure and advantages of stratification.

20. Define simple random sampling and stratified random sampling. What are random numbers and how can you use them?

21. The values of a characteristic x of a population containing six units are given as 2, 6, 5, 1, 7, 3. Take all possible samples of size two and verify that the mean of the population is exactly equal to the mean of the sample means.

22. Distinguish between sampling error and non-sampling error. How can you control non-sampling error?

23. Explain the relative advantages and disadvantages of sampling and census methods for collection of statistical information.

24. Define a randon sample. What is the diffence between random sampling with and without replacement?

Unit 7 ❑ Theory of Estimation

Structure

7.1 Introduction

Statistical inference is that branch of Statistics which is concerned with using probability concept to deal with uncertainty in decision making. The field of statistical inference refers to the process of selecting and using a sample statistic to draw inference about the population parameter based on a sub-set, drawn from the population. The problem of statistical inference can be divided into two categories : 1. Estimation 2. Test of hypothesis.

Estimation : When data are collected by sampling from a population, the most important objective of statistical analysis is to draw inference about that population from the information given in the sample data. Statistical estimation is concerned with

the methods by which population characteristics are estimated from the sample information. The true value of a parameter is an unknown constant that can be correctly ascertained only by an exhaustive study of the population. Statistical estimation procedures provide us with the means of obtaining estimates of population parameters with desired degrees of precision.

Let us consider a random sample $x_1, x_2, \ldots, x_n$ of size n which follows a distribution with an unknown parameter θ. In estimation theory, it is required to find an estimate of θ on the basis of sample values. The estimation of θ can be made in the following two ways : 1) Point estimation and ii) Interval estimation.

Point Estimation : A point estimate is a single number which is used as an estimate of the unknown parameter. Although a point estimate may be the most common way to express an estimate, it suffers from a major limitation since it fails to indicate how close it is to the quantity it is supposed to estimate.

Interval Estimation : An interval estimate of a population parameter is a statement of two values between which it is estimated that the unknown parameter lies with definite probability. An interval estimate would always be specified by two values, i.e. the limits.

7.2 Properties of a Good Estimator

There are four criteria by which we can evaluate the quality of a statistic as an estimator. These are : Unbiasedness, Efficiency, Consistency and Sufficiency.

Unbiasedness : An estimator T is said to be an unbiased estimator of a parameter θ

if $E(T) = \theta \; \forall T$.

If $E(T) \neq \theta$, then T is said to be a biased estimator of θ. The bias of the estimator is given by :

Bias $= E(T) - \theta$.

Example 1 : Let us consider a random sample of size an as $x_1, x_2, \ldots, x_n$ drawn from a population with mean μ and variance s^2. Then we have $E(\bar{x}) = E\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \mu$.

That is, the sample mean, $\bar{x}$, is an unbiased estimator of the population mean μ.

$S^2 = \frac{1}{n}\sum(x_i - \bar{x})^2$ is not an unbiased estimator of σ^2, because $E(S^2) = \frac{n-1}{n}\sigma^2 \neq \sigma^2$

i.e. the sample variance, S^2 is not an unbiased estimator of population variance, σ^2.

Efficiency: An estimator is said to be efficient, if its value remains stable from sample to sample. The best estimator would be that estimator which would have the least variance. Thus if T and T* are two estimators of θ and $V(T) < V(T^*)$, then T is more efficient than T*.

Among three point estimators of central tendency, the arithmetic mean, the median and the mode, the arithmetic mean is considered to have the least variance and hence a better estimator.

An estimator which is unbiased and has also minimum variance is said to be the *minimum variance unbiased Estimator (MVUE).* Also an estimator having minimum variance among all estimators of a population parameter is known as the *most efficient estimator or best estimator.* If an estimator is unbiased and best, then it is known as *the best unbiased estimator.* Further, if the best unbiased estimator is the linear function of the sample observations, it is known as *best linear unbiased estimator (BLUE).* It may be noted that the sample mean is the best linear unbiased estimator of the population mean.

Consistency : It is desirable to have an estimator with a probability distribution that comes closer and closer to the population parameter as the sample size is increased. An estimator possessing this property is called a consistent estimator. An estimator T is said to be consistent estimator of the parameter θ if

$T \to \theta$ with probability 1 as $n \to \infty$.

We may say that the sample mean $\bar{x}$ is a consistent estimator of the population mean μ.

Since $E(\bar{x}) = \mu$ and $V(\bar{x}) = \sigma^2/n \to 0$ as $n \to \infty$.

Sufficiency : An estimator is said to be sufficient if it possesses all the information about the population parameter contained in the sample. If $x, x_2, \ldots x_n$ be a random sample of size n from a population with p.m.f. or p.d.f. $f(x, \theta)$ and T is a sufficient estimator of θ if we have

$f(x, 0) = g(T, 0)\ h(x_1, x_2, \ldots\ldots, x_n)$,

where $g(T, \theta)$ is the sampling distribution of T and $h(x_1, x_2, \ldots\ldots, x_n)$ is independent of θ. Sufficient estimators are the most desirable but are not very commonly available. The following points must to be noted about sufficient estimators :

(i) A sufficient estimator is always consistent.

(ii) A sufficient estimator is not efficient if there exists an efficient estimator.

(iii) A sufficient estimator may or may not be unbiased.

Example 2. If $x_1, x_2, \ldots\ldots, x_n$ is a sample of n independent observations from a normal population with mean μ and variance σ^2, show that $\bar{x}$ is an unbiased estimator of μ but $S^2 = \frac{1}{n}\sum(xi - \bar{x})^2$ is not an unbiased estimator of σ^2.

Solution : We know $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

Now we have to show that $E(\bar{x}) = \mu$ and $E\left[\frac{1}{n}\sum_{i=1}^{n}(xi - \bar{x})^2\right] \neq \sigma^2$.

Since the sample is from a normal population, therefore

$$E(x_i) = \sum_{i=1}^{N} X_i P(x_i = X_i) = \frac{1}{N}\sum_{i=1}^{N} X_i = \mu$$

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

$$= \frac{1}{N}\sum_{i=1}^{N}\left[X_i - E(x_i)\right]^2$$

$$= \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2$$

$$= \sigma^2$$

$$V(\bar{x}) = V\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)$$

$$= \frac{1}{n^2}\sum_{i=1}^{n} V(x_i), \text{ since } x_1, x_2, \ldots\ldots, x_n \text{ are independent.}$$

$$= \frac{1}{n^2} \sum_{i=1}^{n} \sigma^2$$

$$= \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

Now,

$$E(\bar{x}) = E\left(\frac{1}{n} \sum_{i=1}^{n} x_i\right)$$

$$= \frac{1}{n} \sum_{i=1}^{n} E(x_i)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \mu$$

$$= \frac{1}{n} n\mu = \mu$$

Again,

$$E\left(\frac{1}{n} \sum x_i^2 - \bar{x}^2\right)$$

$$= E\left(\frac{1}{n} \sum_{i=1}^{n} x_i^2 - \bar{x}^2\right)$$

$$= \frac{1}{n} \sum_{i=1}^{n} E(x_i^2) - E(\bar{x}^2)$$

$$= \frac{1}{n} \sum_{i=1}^{n} [V(x_i) + E^2(x_i)] - [V(\bar{x}) + E^2(\bar{x})]$$

$$= \frac{1}{n} \sum_{i=1}^{n} (\sigma^2 + \mu^2) - \left(\frac{\sigma^2}{n} + \mu^2\right)$$

$$= \frac{1}{n} n\sigma^2 + \frac{1}{n} n\mu^2 - \frac{\sigma^2}{n} - \mu^2$$

$$= \sigma^2 - \frac{\sigma^2}{n}$$

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

$$\neq \sigma^2$$

Hence, $\bar{x}$ is an unbiased estimator of μ but $\frac{1}{n}\sum(x_i - \bar{x})^2$ is not an unbiased cestimator of σ^2.

Note : $E\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right) = \frac{n-1}{n}\sigma^2$

or, $E\left(\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right) = \sigma^2$

Therefore, $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ is an unbiased estimator of σ^2.

Example 3 : If T_1, T_2 and T_3 are independent unbiased estimates of θ and all have the same variance, which of the following unbiased estimates of θ would you prefer?

$$\frac{T_1 + 2T_2 + T_3}{4}, \frac{3T_1 + T_2 - 2T_3}{2}, \frac{T_1 + T_2 + T_3}{3}$$

Solution : Here $V(T_1) = V(T_2) = V(T_3) = \sigma^2$.

Now, $V\left(\frac{T_1 + 2T_2 + T_3}{4}\right)$

$$= \frac{1}{16}\left[V(T_1) + 4V(T_2) + V(T_3)\right]$$

$$= \frac{1}{16}\left[\sigma^2 + 4\sigma^2 + \sigma^2\right]$$

$$= \frac{6}{16}\sigma^2 = \frac{3}{8}\sigma^2.$$

$$V\left(\frac{3T_1 + T_2 - 2T_3}{2}\right)$$

$$= \frac{1}{4}\left[9V(T_1) + V(T_2) + 4V(T_3)\right]$$

$$= \frac{1}{4}[9\sigma^2 + \sigma^2 + 4\sigma^2]$$

$$= \frac{14}{4}\sigma^2 = \frac{7}{2}\sigma^2.$$

Finally,

$$V\left(\frac{T_1 + T_2 + T_3}{3}\right)$$

$$= \frac{1}{9}\left[V(T_1) + V(T_2) + V(T_3)\right]$$

$$= \frac{3}{9}\sigma^2 = \frac{1}{3}\sigma^2$$

Clearly $V\left(\frac{T_1 + T_2 + T_3}{3}\right)$ is minimum

So, we say that $\left(\frac{T_1 + T_2 + T_3}{3}\right)$ would be the most preferable.

7.3 Methods of Point Estimation

For obtaining point estimates of parameters of a probability distribution, several methods are available. These are :

i) Method of moments

ii) Maximum likelihood Method

iii) Least squares method

iv) Minimum chi-square method

We shall discuss here the method of moments and the maximum likelihood Method only.

7.3.1 Method of moments

This is one of the classical methods and the motivation comes from the fact that the sample moments are in some sense estimates for the population moments. Thus according to this principle the sample moments are equated to the population moments and by using these equations the parameters in a given population are estimated. That is, the parameters are estimated by using the relations, $m'_r = \mu'_r$, r = 1, 2, 3,..... where m'_r and μ'_r are sample and population moments about the origin respectively. We can equate corresponding central moments as well. For simplicity, moments of lower order are usually taken.

Example 4 : Suppose we are to estimate the parameter X of a Poisson distribution on the basis of sample observations $x_1, x_2, \ldots\ldots\ldots\ldots x_n$. We know that, for the Poisson distribution $E(\bar{x}) = \lambda = \mu'_1$ and the sample mean $E(\bar{x}) = E\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = m'_1$

Writing $m'_1 = \mu'_1$ we get an estimate of λ as $\hat{\lambda} = \bar{x}$ = the sample mean.

7.3.2 Maximum likelihood Method

This method is simple and it gives estimates that possesses many desirable properties. Let x_1, x_2, x_n be a random sample from a population with probability mass function or probability density function $f(x; \theta)$, involving a parameter θ. For fixed θ, the function $f(x_1, x_2 \ldots\ldots, x_n ; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$ may be looked upon as a function of the sample observations. But when $x_1, x_2, \ldots\ldots x_n$ are given, the above function may be looked upon as a function of θ, which is called the likelihood function of θ and is denoted by $L(\theta)$. The principle of maximum likelihood suggests to take that value as an estimate of θ for which $L(\theta)$ is a maximum. That is, we may choose $\hat{\theta}$ such

that $\left[\frac{dL(\theta)}{d\theta}\right]_{\theta=\hat{\theta}} = 0$ and $\left[\frac{d^2L(\theta)}{d\theta^2}\right]_{\theta=\hat{\theta}} < 0$.

Since log L(θ) is a monotonic increasing function of L(θ), log L (θ) is also maximum for the same value for θ, and it is more generally worked with log L(0).

We should note that

(a) When the derivative does not exist at θ the method fails.

(b) The technique of differentiation is a convenient tool but it is not required in all the problems.

(c) If there are a number of maxima in a particular problem, then the one corresponding to the largest ordinate is taken as the maximum liklihood estimate.

Example 5 : For a normal population with parameters μ and σ^2, obtain the maximum likelihood estimators of the parameters.

Solution : The p.d.f. of the normal distribution $f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$.

For a random sample of n independent observations the likelihood function L is given by

$$L(\mu, \sigma) = \prod_{i=1}^{n} f(x_i; \mu, \sigma) = \frac{1}{(\sigma\sqrt{2\pi})^n} e^{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2}$$

Taking logarithm on both sides,

$$\log L = -n \log \sigma - \frac{n}{2} \log 2\pi - \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2$$

(i) MLE of μ

$$\frac{\partial \log L}{\partial \mu} = \frac{1}{2} 2\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)\left(-\frac{1}{\sigma}\right) = 0.$$

or, $$\sum_{i=1}^{n}(x_i-\mu) = 0.$$

or, $\sum_{i=1}^{n} x_i - n\mu = 0$

or, $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i - \bar{x}$

or $\mu = \bar{x}$ = the sample mean.

(ii) MLE of σ^2

Let us differentiate logL with respect to σ to get

or, $\frac{\partial \log L}{\partial \sigma} = \frac{n}{\sigma} - \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\left(-2\frac{1}{\sigma^3}\right)$

or, $-\frac{n}{\sigma} + \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{\sigma^3} = 0$

or, $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$.

Hence, the MLE of μ is $\bar{x}$ and σ^2 is $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

7.4 Interval Estimation

In the theory of interval estimation, it is desired to find an interval, based on sample values, which is expected to include the unknown parameter with a specified probability. Let $x_1, x_2, \ldots\ldots, x_n$ be a random sample from a population with an unknown parameter θ. We will try to find two functions t_1 and t_2 of the sample values such that the probability of θ being included in the random interval (t_1, t_2) has a given value, say $1 - \alpha$. So

$P(t_1 \leq 0 \leq t_2) = 1 - \alpha$.

Here the interval (t_1, t_2) is called a $100(1 - \alpha)\%$ confidence interval for the parameter 0. The quantities t_1 and t_2 which serve as the lower and upper limits of the

interval are known as confidence limits. $(1 - \alpha)$ is called the confidence coefficient This has the meaning that if samples of same size n are taken and if the interval (t_1, t_2) is constructed for every sample, then in the long run $100 (1 - \alpha)\%$ of the intervals will cover the unknown parameter θ. It may be noted that the probability statement is on the random interval (t_1, t_2) rather than on the parameter θ.

It is to be noted that a $100 (1 - \alpha)\%$ confidence interval is not unique. We look for that interval which is shorter than any other interval with the same confidence coefficient. If a confidence interval is constructed by omitting equal tail areas, then we get what is known as central interval. In a symmetrical distribution, it can be shown that the central interval is the shortest.

In practice, the method of finding confidence interval consists in first finding a random variable, call it v, that involves the sample values and the desirable parameter θ but whose distribution does not depend on any unknown parameter. Next two numbers v_1 and v_2 are chosen such that

$$P(v_1 \leq v \leq v_2) = 1 - \alpha,$$

where $(1 - \alpha)$ is the desired confidence coefficient, such as 0.95, 0.99 etc. Then this inequality is solved so that the probability statement assumes the form

$$P(t_1 \leq \theta \leq t_2) = 1 - \alpha,$$

where t_1 and t_2 are random variables depending on v but not involving θ. Finally, one substitutes the sample values in t_1 and t_2 to obtain a numerical interval which is then the desired confidence interval.

As compared to point estimation, interval estimation is better as it takes into account the variability of the estimator in addition to its single value and thus, provides a range of values. Unlike point estimation, interval estimation indicates that estimation is an uncertain process.

The methods of construction of confidence intervals in various situations are explained through the following examples.

7.4.1 Confidence Interval for Population Mean

a) $100 (1 - \alpha)\%$ confidence interval for μ in a random sample of size n from a normal population with mean μ and standard deviation σ is as follows :

$$\left(\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

where Z_α is the upper α point of the standard normal distribution. That is, $P(Z \geq Z_\alpha) = \alpha$.

Note : If the sample size is sufficiently large, then for unknown standard deviation σ, one can obtain the interval by using the standard normal distribution.

(b) If the standard deviation σ is unknown, then for small sample, the above interval will be obtained by using the t-distribution which is as follows :

$$\left(\bar{X} - t_{\alpha/2;\, n-1}\frac{s}{\sqrt{n}}, \bar{X} + t_{\alpha/2;\, n-1}\frac{s}{\sqrt{n}}\right),$$

where $t_{\alpha\,;\,n-1}$ is the upper α–point of the t-distribution with (n–1) degrees of freedom.

Example 6 : Construct 95% confidence intervals for mean of a normal population.

Solution : Let $x_1, x_2, \ldots\ldots, x_n$ be a random sample of size n from a normal population with mean μ and standard deviation σ.

We know that sampling distribution of $\bar{X}$ is normal with mean μ and standard error $\frac{\sigma}{\sqrt{n}}$. Therefore, $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ will be the standard normal variate.

From the tables of areas under the sandard normal curve, we can write

$P[-1.96 \leq Z \leq 1.96] = 0.95$

or, $$P\left[-1.96 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq 1.96\right] = 0.95 \qquad \ldots\ldots (7.1)$$

The inequality $-1.96 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ can be written as

$$-1.96\,\frac{\sigma}{\sqrt{n}} \leq \bar{X} - \mu \text{ or } \mu \leq \bar{X} + 1.96\frac{\sigma}{\sqrt{n}} \qquad \ldots\ldots (7.2)$$

Similary, from the inequality $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq 1.96$, we can write

$$\mu \geq \bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \qquad \text{..... (7.3)}$$

Combining (7.2) and (7.3), we get

$$\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}.$$

Thus, we can write equation (7.1) as

$$P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96\frac{\sigma}{m}\right) = 0.95.$$

This gives us a 95% confidence interval for the parameter μ. The lower limit of μ is $\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}$ and the upper limit is $\bar{X} + 1.96\frac{\sigma}{\sqrt{n}}$. The probability of μ lying between these limits is 0.95 and, therefore, this interval is also termed as 95% confidence interval for μ.

In a similar way, we can construct a 99% confidence interval for μ as

$$P\left(\bar{X} - 2.58\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 2.58\frac{\sigma}{\sqrt{n}}\right) = 0.99.$$

Thus, the 99% confidence limits for μ are $\bar{X} \pm 2.58\frac{\sigma}{\sqrt{n}}$.

Remarks : When σ is unknown and $n < 30$, we use t value instead of 1.96 or 2.58 and use s in place of σ.

7.4.2 Confidence Interval for Population Proportion

100 $(1 - \alpha)$% confidence interval for π from a binomial population with parameters n and π with sufficiently large n is as follows :

$$\left(p - Z_{\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}}, p + Z_{\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}}\right),$$

where Z_α is the upper α - point of the standard normal distribution.

Since π is not known in general, its estimator p is used in the estimation of

standard error of p, i.e. estimated S. E. (p) = $\sqrt{\frac{p(1-p)}{n}}$. Then the corresponding confidence interval will be

$$\left(p - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}, p + Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right).$$

Example 7 : Obtain the 95% confidence limits for the proportion of successes in a binomial population.

Solution : Let the parameter π denote the proportion of successes in a binomial population. Further, let p denote the sample proportion of successes in n ($\geq$ 50) trials. We know that the sampling distribution of p will be approximately normal with mean π and standard error $\sqrt{\frac{\pi(1-\pi)}{n}}$.

Since π is not known, therefore, its estimator p is used in the estimation of the standard error of p, i.e. estimated S.E.(p) = $\sqrt{\frac{p(1-p)}{n}}$.

Thus, the 95% confidence interval for p is given by

$$P\left(p - 1.96\sqrt{\frac{p(1-p)}{n}} \leq \pi \leq p + 1.96\sqrt{\frac{p(1-p)}{n}}\right) = 0.95.$$

This gives the 95% confidence limits as $p \pm 1.96\sqrt{\frac{p(1-p)}{n}}$.

Example 8 : A sample of 1600 screws is taken from a large consignment. 64% of the screws were found to be defective. Assuming that the simple sampling conditions hold good, estimate the confidence limits of the proportion of defective screws.

Solution : Let π be the proportion of defective screws. Corresponding proportion in the sample is given as p = 0.64.

$$\therefore \text{ S.E. (p)} = \sqrt{\frac{0.64 \times 0.36}{1600}} = \frac{0.48}{40} = 0.012$$

We know from the normal distribution that almost whole of the distribution lies

between 3σ limits. Therefore, the 99% confidence interval is given by $P[p-2.58S.E.(p) \leq \pi \leq p+2.58S.E.(p)] = 0.9973$.

Thus, the 99% confidence limits are 0.609 $(=0.64 - 2.58 \times 0.012)$ and 0.671 $(=0.64 + 2.58 \times 0.012)$ respectively.

Hence, the proportions of defective screws in large consignment are beween 60.9% and 67.1%.

Example 9 : A random sample of 100 items taken from a large batch of articles contains 5% defective items. (a) Set up 96% confidence limit for the proportion of defective items in a batch. (b) If the batch contains 2,697 items, set up 95% confidence limits for the proportion of defective items.

Solution : Here $n = 100$.

p = proportion of defective in the sample $= 5/100 = 0.05$.

(a) Here estimate of S. E. of p is given by

$$S.E.(p) = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.05 \times 0.95}{100}} = 0.02179 \cong 0.022 .$$

From the table of the normal distribution

$$Z_{0.02} = 2.05.$$

Hence 96% confidence limits for the population proportion of defectives are :

$$p \pm Z_{0.02}\ S.E.\ (p) = 0.05 \pm 2.05 \times 0.022 = (0.005,\ 0.095).$$

(b) If $N = 2669$, then 95% confidence limits for p is given by

$$p \pm Z_{0.025}\ SE(p) = p \pm 1.96 \times \sqrt{\frac{(N-n)pq}{N(n-1)}}$$

$$\left[\sqrt{\frac{pq}{n}\frac{(N-n)}{N-1}} \text{ is not unbiased but } \sqrt{\frac{pq}{n-1}\frac{N-n}{N}} \text{ is unbiased } \right]$$

$$= 0.050 \pm 1.96 \times \sqrt{\frac{(2669-100)}{2669} \times \frac{0.05 \times 0.95}{99}}$$

$$= 0.050 \pm 1.96 \times 0.0215 = 0.050 \pm 0.042$$

$$= (0.008,\ 0.092).$$

7.5 Determination of an Approximate Sample Size for a specified Confidence Level

Le us assume that we want to find the size of a sample to be taken from the population such that the difference between sample mean and population mean would not exceed a given value, say E, with a given level of confidence. In other words, we want to find n such that

$$P(|\bar{X}-\mu| \le E) = 1-\alpha, \text{ say} \qquad \ldots\ldots (7.4)$$

Assume that the sampling distribution of $\bar{X}$ is normal with a mean μ and standard error $\frac{\sigma}{\sqrt{n}}$, Therefore, we can write $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ which will be a standard normal variate. If Z_α be the upper α - point of the standard normal distribution, then from the tables of areas under the standard normal curve, we can write.

$$P\left(-Z_{\alpha/2} \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le Z_{\alpha/2}\right) = 1-\alpha$$

or, $$P\left(\left|\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right| \le Z_{\alpha/2}\right) = 1-\alpha$$

or, $$P\left(|\bar{X}-\mu| \le Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha \ldots.(7.5)$$

Comparing (7.4) and (7.5), we get

$$Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = E \quad \text{or} \quad n = Z^2_{\alpha/2}\frac{\sigma^2}{E^2}.$$

The lesser the magnitude of E, the more precise will be the interval estimate.

7.6 Exerciese

1. Distinguish between point estimation and inteval estimation.
2. What are the criteria of a good estimator?

3. What is a consistent estimator?

4. Define a sufficient statistic.

5. Briefly discuss the importance of estimation theory in decision making in the face of uncertainty.

6. What do you mean by unbiased estimator and minimum variance unbiased estimator?

7. If $x_1, x_2, \ldots\ldots, x_n$ be a random sample from $N(\mu, \sigma^2)$, obtain the maximum likelihood estimator of μ and σ and check whether these are unbiased or not.

8. Verify that the estimates of Posisson parameter obtained by the method of moments and by the maximum likelihood method are identical.

9. Explain the concept of confidence interval, confidence limits and confidence coefficient.

10. If the variance of two unbiased estimators T_1 or T_2 of θ are same, which of $\frac{T_1+T_2}{2}$ and $\frac{3T_1+T_2}{4}$ is the minimum variance unbiased estimator of θ ?

11. If T_1, T_2 and T_3 are 3 statistics with expectations $E(T_1) = 3\theta_1 + 2\theta_2 + \theta_3$, $E(T_2) = 2\theta_1 + 3\theta_2 + \theta_3$ and $E(T_3) = \theta_1 + \theta_2 + \theta_3$, find the unbiased estimator of θ_1, θ_2 and θ_3.

12. What do you mean by best linear unbiased estimator? Give an example.

13. A random sample of 100 days shows an avarage daily sales of Rs. 5,000/- with an s.d. of Rs. 1000/- in a particular shop. Assuming a normal distribution, construct the 95% confidence interval for the mean sales per day.

14. A random sample of size 1000 selected from a large lot of machine parts shows that 7% are defective. What information can be inferred about the percentage of defective in the lot?

15. On the basis of a random sample of size 10 from a normal population with mean 50 and variance 144, find 99% confidence limits for the population mean.

16. Find the sample size such that the probability of the sample mean differing from the population mean by not more than $1/10^{th}$ of standard deviation is 0.95.

17. With sample size of 635 the calculated standard error of mean is 3 with a mean of 150. What sample size should be taken so that we could be 95% confident that the population mean lies within ± 3.5 of the sample mean.

18. In a market area there are 300 shops. A researcher wants to estimate the number of customers visiting these shops per day. The researcher also wants that the sampling error in these estimates is not larger than ± 10 with 95% confidence. The previous studies indicate that the s.d. of the customer arrivals is 85. If the cost per interview is Rs.20, calculate the total cost of the survey involved.

19. In a marketing survey for the introduction of a new product in a town, a sample of 600 persons was drawn. When they were approached for sale, 180 of them purchased the product. Find 95% confidence limits for the percentage of persons who would buy the product in the town.

20. A researcher wishes to estimate the mean of a population by using sufficiently large sample. The probability is 0.95 that the sample mean will not differ from the true mean by more than 50% of the standard deviation. How large a sample should be taken?

Unit 8 ❑ Test of Hypothesis

Structure

8.1 Introduction

A test of statistical hypothesis is a statistical procedure which, when the sample values have been obtained, leads to a decision to accept or to reject the hypothesis under consideration. In many cases we are to make decisions about populations on the basis of sample data. Some information as to the feature of the population or the hypothetical values of the parameters may be available and on the basis of certain rules or criteria we may decide whether the hypothesis is acceptable or not in the light of the sample data collected from the population. This is the problem of hypothesis testing or test of signficance. The theory of hypothesis testing begins with a basic assumption about the parameter of the population. This assumption is termed as hypothesis made on the basis of the sample observations. The validity of a hypothesis will be tested by analysing the sample. The procedure which enables us to decide whether a certain hypothesis is true or not, is called test of hypothesis.

8.1.1 Some Definitions

Statistical hypothesis

A statistical hypothesis is an assertion about the probability distribution of random variables which is verified on the basis of a sample.

Null hypothesis and alternative hypothesis

The null hypothesis is that which is tested for possible rejection under the assumption that it is true. It is denoted by H_0. This hypothesis asserts that there is no difference between population and sample in the matter under consideration.

Any hypothesis which contradicts the null hypothesis is called an alternative hypothesis. It is denoted by H_1.

For example [$H_0 : \mu = \mu_0$ against alternatives a) $H_1: \mu > \mu_0$

or b) $H_1 : \mu < \mu_0$

or c) $H_1: \mu \neq \mu_0$

Test Statistic

Any statistic is a function of sample observations. Test statistic is a statistic whose computed value determines the final decisions regarding acceptance or rejection of

the null hypothesis. The appropriate test statistic is to be chosen very carefully and knowledge of its sampling distribution under null hypothesis is essential in framing decision rules. If the value of the test statistic falls in the critical region, the null hypothesis is rejected.

Level of significance

This is the probability level, under the null hypothesis, which is employed in defining the critical region. It is generally denoted by the symbol α and is usually taken to be 0.05 or 0.01.

Critical region and acceptance region

The set of values of the test statistic which leads to the rejection of the null hypothesis is known as *critical region or rejection region* of the test. On .the other hand, the values that lead to the acceptance of the null hypothesis are said to form the *acceptance region.* Here we are to test the validity of H_0 against that of H, at a certain level of significance. In a normal distribution the area under the normal curve outside the ordinates at mean $\pm$ 1.96 (s.d.) is only *5%,* the probability that the observed value of the statistic differs from the expected value of 1.96 times the standard error or more is 0.05, and the probability of a larger difference will be still smaller.

Let $Z = \frac{\sqrt{n}(\bar{x} - \mu_0)}{\sigma}$ and then if $|Z| \geq 1.96$, we reject the null hypothesis.

Therefore $|Z| \geq 1.96$ constitutes the critical region of the test. It is denoted by ω, Thus $|Z| \leq 1.96$ constitutes the acceptance region of the test and it is denoted by ω.

Type I and Type II error

Probability of type I error is defined as the probability of rejecting the null hypothesis when it is true. The critical region is so determined that the probability of type I error does not exceed the level of significance of the test.

Probability of Type I error = $P(x \varepsilon \omega / H_0)$

Probability of type II error is defined as the probability of accepting the null hypothesis when it is really false.

Probability of type II error = $P(x \varepsilon \omega - \omega / H_1) = 1 - P(x \varepsilon \omega / H_1)$.

Power of a test

The power of a test is defined as the probability of rejecting the null hypothesis

when it is false. On the other hand, power of a test is defined as

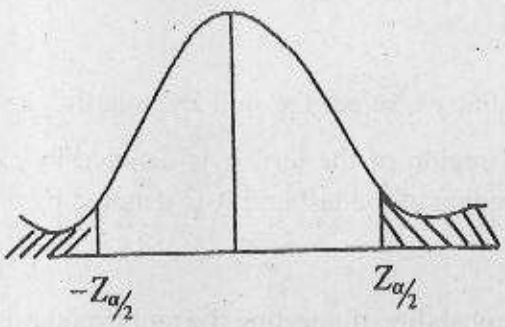
Power = 1 - Probability of type II error

$= P(x \varepsilon \omega / H_1)$

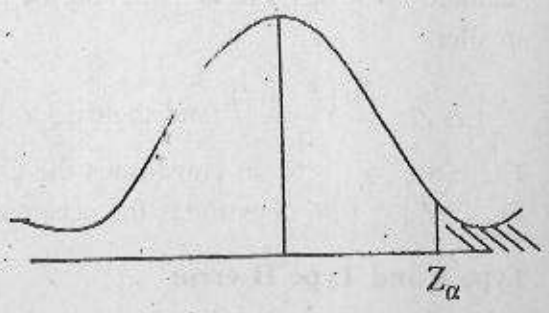
Two-tailed and one-tailed test

The specification of a critical region for a test depends upon the nature of the alternative hypothesis and the value of α. For example, $H_1 : \mu \neq \mu_0$, this implies that μ, may be less or greater than μ_0. Thus, the critical region is to be specified on "both tails of the curve with each part corresponding to half of the value of α. A test having critical region at both the tails of the probability curve is termed as a two ailed test.

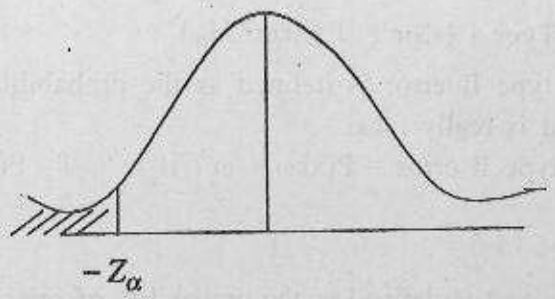
Further, if $H_1 : \mu > \mu_0$ or $\mu < \mu_0$ the critical region is to be specified only at one tail of the probability curve and the corresponding test is termed as a one-tailed test.



Critical region for two-tailed test.



Critical region for one-tailed (right tail) test



Critical region for one-tailed (left tail) test

8.2 Tests of Hypothesis Concerning Mean of Single Population

These tests can be divided into two broad categories depending upon whether σ, the population standard deviation, is known or not.

8.2.1 Test of Hypothesis Concerning Specified Mean (σ being known)

This test is applicable when the random sample $X_1, X_2, \ldots\ldots, X_n$ is drawn from a normal population with mean μ and standard deviation σ. We can consider the test in the folowing steps :

Step I. (Hypothesis Formulation)

We set up the null hypothesis and alternative hypothesis on the basis of the given problem. That is,

$H_0 : \mu = \mu_0$ (specified) against alternatives a) $H_1 : \mu > \mu_0$

or b) $H_1 : \mu < \mu_0$

or c) $H_1 : \mu \neq \mu_0$

Step II. (Test Statistic)

To test the above null hypothesis we consider the appropriate test statistic

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma} \sim N(0, 1) \text{ under the null hypothesis.}$$

Step III. (Computation)

$$Z_{Cal} = \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma} \text{ (say)}.$$

Step IV. (Conclusion)

(a) Reject the null hypothesis $H_0 : \mu = \mu_0$ against the alternative $H_1 : \mu > \mu_0$ at α level of significance if $Z_{Cal} > Z_\alpha$.

b) Reject the null hypothesis $H_0 : \mu = \mu_0$ against the alternative $H_1 : \mu < \mu_0$ at α level of significance if $Z_{Cal} < - Z_\alpha$ and

c) Reject the null hypothesis $H_0 : \mu = \mu_0$ against the alternative $H_1 : \mu \neq \mu_0$ at α level of significance if $| Z_{Cal} | > Z_{\alpha/2}$

Here Z_α denotes the upper α-point of a standard normal distribution. That is, $P(Z > Z_\alpha) = \alpha$, where Z is a standard normal variate.

Example 1. Suppose a beverage company wished to study whether the average per household consumption of tea is more than 200 gms with a possible variation of 10 gms. The researcher collected 10 samples of households and found that the consumptions (gms) were :

175, 225, 190, 210, 200, 180, 220, 230, 150 and 160. Do the data indicate that the consumption is below 200 gms?

Solution. Let X be the random variable denoting consumption of tea of a household. We assume that X follows the normal distribution with a mean μ and standard deviation σ. Let $X_1, X_2, ..., X_n$ be a random sanple of size n from a normal population with mean μ and standard deviation σ.

Here we have to test

$H_0 : \mu = 200$ against the alternative $H_1 = \mu < 200$.

The appropriate test statistic is

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma} \sim N(0, 1) \text{ under the null hypothesis.}$$

$$Z_{Cal} = \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma} = \frac{195 - 200}{10/\sqrt{10}} = -1.7.$$

Since $Z_{Cal} = -1.7 < -Z_{0.05} = -1.64$, so the H_0 is rejected at 5% level of significance. Hence the average consumption per household can be taken as below 200 gms.

Example 2. Construct 95% confidence interval for mean of a normal population.

Solution : Let $X_1, X_2 ... X_n$ be a random sample of size n from a normal population with

mean μ and standard deviation σ.

We know that sampling distribution of $\bar{X}$ is normal with mean μ and standard error $\frac{\sigma}{\sqrt{n}}$. Therefore, $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ will be a standard normal variate.

From the tables of areas under the standard normal curve, we can write

$$P[-1.96 \le Z \le 1.96] = 0.95 \text{ or } P\left[-1.96 \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le 1.96\right] = 0.95 \ldots (8.1)$$

The inequality $1.96 \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ can be written as

$$-1.96\frac{\sigma}{\sqrt{n}} \le \bar{X}-\mu \text{ or } \mu \le \bar{X}+1.96\frac{\sigma}{\sqrt{n}}. \quad \ldots\ldots (8.2)$$

Similarly, from the inequality $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le 1.96$ we can write

$$\mu \le \bar{X}-1.96\frac{\sigma}{\sqrt{n}} \quad \ldots (8.3)$$

Combining (8.2) and (8.3), we get

$$\bar{X}-1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X}+1.96\frac{\sigma}{\sqrt{n}}$$

Thus, we can write equation (8.1) as

$$P\left(\bar{X}-1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X}+1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95.$$

This gives us a 95% confidence interval for the parameter μ. The lower limit of μ is $\bar{X}-1.96\frac{\sigma}{\sqrt{n}}$ and the upper limit is $\bar{X}+1.96\frac{\sigma}{\sqrt{n}}$. The probability of μ lying between these limits is 0.95 and, therefore, this interval is also termed as 95% confidence interval for μ.

In a similar way, we can construct a 99% confidence interval for μ as

$$P\left(\bar{X}-2.58\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X}+2.58\frac{\sigma}{\sqrt{n}}\right) = 0.99.$$

Thus, the 99% confidence limits for μ are $\bar{X} \pm 2.58\frac{\sigma}{\sqrt{n}}$.

Remark : When σ is unknown and $n < 30$, we use t value instead of 1.96 or 2.58 and use s in place of σ.

8.2.2 Test of Hypothesis Concerning Specified Mean (σ being unknown)

This test is applicable when the random sample $X_1, X_2, ..., X_n$ is drawn from a normal population with mean μ and standard deviation σ. We can consider the test in the following steps :

Step 1. (Hypothesis Formulation)

The null hypothesis and the alternative hypothesis on the basis of the given problem are as follows :

$H_o : \mu = \mu_0$ (spcified) against alternatives a) $H_1 : \mu > \mu_0$

or b) $H_1 : \mu < \mu_0$

or c) $H_1 : \mu \neq \mu_0$

Step II. (Test Statistic)

To test the above null hypothesis we consider the appropriate test statistic

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} = \frac{\sqrt{n}(\bar{X} - \mu_0)}{s} \sim t - \text{ distribution with (n 1) degrees of freedom}$$

under null hypothesis, where $s = \sqrt{\frac{\Sigma(X_i - \bar{X})^2}{n-1}}$.

Step III. (Computation)

Let the value of this statistic calculated from sample be denoted as $t_{Cal} = \frac{\sqrt{n}(\bar{X} - \mu_0)}{s}$ (say).

Step IV. (Conclusion)

(a) Reject the null hypothesis $H_0 : \mu = \mu_0$ against the alternative $H_1 : \mu > \mu_0$ at α level of significance if $t_{Cal} > t_{\alpha, n-1}$.

(b) Reject the null hypothsis $H_0 : \mu = \mu_0$ against the alternative $H_1 : \mu < \mu_0$ at α level of significance if $t_{Cal} < - t_{\alpha, n-1}$ and

(c) Reject the null hypothesis $H_0 : \mu = \mu_0$ against the alternative $H_1 : \mu \neq \mu_0$ α level of significance if $|t_{Cal}| > t_{\alpha/2,\ n-1}$.

Here $t_{\alpha,\ n-1}$ denotes the upper α-point of a-t-distribution with $n - 1$ degrees of freedom. That is,

$P(t > t_{\alpha,\ n-1}) = \alpha$.

Note : When s is not known, we use its estimate computed from the given samle. Here, the nature of the sampling distribution of $\bar{X}$ would depend upon sample size n. There are the following two possibilities :

(i) If the parent population is normal and $n \leq 30$ (popularly known as small sample case), use t - test. The unbiased estimate of σ in this case is given by

$$s = \sqrt{\frac{\Sigma(X_i - \bar{X})^2}{n-1}}.$$

(ii) If $n > 30$ (large sample case), use the standard normal test. The unbiased estimate of σ in this case can be taken as $s = \sqrt{\frac{\Sigma(X_i - \bar{X})^2}{n-1}}$, since the difference between n and $n - 1$ is negligible for large values of n. Note that the parent population may or may not be normal in this case.

Example 3. A company making ice-cream and sells it in 500 gms packs. Periodically, a sample of 16 packs is taken and sample mean is found to be 460 gms. and the unbiased estimated standard deviation of 40 gms. Does the sample mean differ significantly from the intended weight of 500 gms?

Solution : Let X be the random variable denoting weight of an ice-cream pack. We assume that X follows the normal distribution with a mean μ and sandard deviation σ. $X_1, X_2, \ldots\ldots, X_n$ be a random sample of size n from a normal population with mean μ and standard deviation σ.

Here we have to test

$H_0 : \mu = 500$ against the alternative $H_1 = \mu \neq 500$.

To test the above null hypothesis we consider the appropriate test statistic $t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} = \frac{\sqrt{n}(\bar{X} - \mu_0)}{s} \sim t$ distribution with $(n - 1)$ degrees of freedom under

the null hypothesis, where $s = \sqrt{\frac{\sum(X_i - \bar{X})^2}{n-1}}$

$$t_{Cal} = \frac{\sqrt{n}(\bar{X} - \mu_0)}{s} = \frac{\sqrt{16}(460 - 500)}{40} = -4.$$

Since $|t_{Cal}| = 4 > t_{0.025} = 2.131$, so the H_0 is rejected at 5% level of significance. Hence the sample mean differs significantly from the intended weight of 500 gms.

8.3 Tests of Hypothesis Concerning Means of Two Populations

These tests can be divided into several categories depending upon whether σ_1 and σ_2, the population standard deviations are known or not. The populations may be dependent or independent.

8.3.1 Test of Hypothesis Concerning Equality of Two Means (σ_1 and σ_2 being known)

This test is applicable when two random samples of sizes n_1 or n_2 drawn from two independent normal populations with unknown means μ_1 and μ_2 and standard deviations σ_1 and σ_2 . We can consider the test in the following steps :

Step I. (Hypothesis Formulation)

We set up the null hypotesis and the alternative hypothesis on the basis of the above problem. That is,

$H_0 : \mu_1 = \mu_2$ against alternatives

a) $H_1 : \mu > \mu_1$

or b) $H_1 : \mu < \mu_2$

or c) $H_1 : \mu \neq \mu_2$

Step II. (Test Statistic)

To test the above null hypothesis we consider the appropriate test statistic

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$ under the null hypothesis.

Setp III. (Computation)

Let the value of this statistic calculated from the sample be denoted as

$$Z_{Cal} = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \text{ (say).}$$

Step IV. (Conclusion)

(a) Reject the null hypothesis $H_0 : \mu_1 = \mu_2$ against the alternative $H_1 : \mu_1 > \mu_2$ at α level of significance if $Z_{Cal} > Z_\alpha$.

(b) Reject the null hypothsis $H_0 : \mu_1 = \mu_2$ against the alternative $H_1 : \mu_1 < \mu_2$ at α level of significance if $Z_{Cal} > - Z_\alpha$ and

(c) Reject the null hypothesis $H_0 : \mu_1 = \mu_2$ against the alternative $H_1 : \mu_1 \neq \mu_2$ at α level of significance if $|Z_{Cal}| > Z_{\alpha/2}$.

Here Z_α denotes the upper α - point of a standard normal distribution. That is, $P(Z > Z_\alpha) = \alpha$, where Z is a standard normal variate.

8.3.2. Test of Hypothesis Concerning Equality of Two Means (σ_1 and σ_2 unknown but $\sigma_1 = \sigma_2$)

This test is applicable when two random samples of sizes n_1 or n_2 drawn from two independent normal populations with unknown means μ_1, μ_2 and standard deviations σ_1 and σ_2 The population standard deviations are assumed to be equal.

Step I. (Hypothesis Formulation)

We set up the null hypothesis and alternative hypothesis on the basis of the above problem. That is,

$H_0 : \mu_1 = \mu_2$ against alternatives

a) $H_1 : \mu_1 > \mu_2$

or b) $H_1 : \mu_1 < \mu_2$

or c) $H_1 : \mu_1 \neq \mu_2$

Step II. (Test Statistic)

To test the above null hypothesis we consider the appropriate test statistic

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_2 - \mu_2)}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{\bar{X}_1 - \bar{X}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$ under the null hypothesis,

where the pooled estimate of σ, denoted by s, is defined as

$$s = \sqrt{\frac{\Sigma(X_{1i} - \bar{X}_1)^2 + \sum(X_{2i} - \bar{X}_2)^2}{n_1 + n_2 - 2}} = \sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}}$$

$$= \sqrt{\frac{(n_1 - 1)s_1^2 + (n_1 - 1)s_1^2}{n_1 + n_2 - 2}}.$$

Setp III. (Computation)

Let the value of this statistic calculated from the sample be denoted as

$$t_{Cal} = \frac{(\bar{X}_1 - \bar{X}_2)}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \text{ (say).}$$

Step IV. (Conclusion)

(a) Reject the null hypothesis $H_0 : \mu_1 = \mu_2$ against the alternative $H_1 : \mu_1 > \mu_2$ at α level of significance if $t_{Cal} > t_{\alpha ; (n_1+n_2-2)}$.

(b) Reject the null hypothsis $H_0 : \mu_1 = \mu_2$ against the alternative $H_1 : \mu_1 < \mu_2$ at α level of significance if $t_{Cal} < t_{\alpha;(n_1+n_2-2)}$ and

(c) Reject the null hypothesis $H_0 : \mu_1 = \mu_2$ against the alternative $H_1 : \mu_1 \neq \mu_2$ at α level of significance if $| t_{Cal} | > t_{\alpha/2;(n_1+n_2-2)}$.

Here $t_{\alpha;(n_1+n_2-2)}$ denotes the upper α - point of a t-distribution with n_1+n_2-2) degrees of freedom. That is, $P(t > t_{\alpha;(n_1+n_2-2)} = \alpha$.

Note : If we consider two random samples of sizes n_1 or n_2 drawn from two independent normal populations with unknown means μ_1, μ_2 and standard deviations σ_1, σ_2 respectively, then

$$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2 \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right).$$

Case I : If σ_1 and σ_2 are known, we use the standard normal test.

(a) To test $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$ (two-tailed test) the test statistic is

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1) \text{ under } H_0 .$$

This value is compared with 1.96 (2.58) for 5% (1%) level of significance.

(b) To test $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 > \mu_2$ (one-tailed test), the test statistic is

$$Z = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1) \text{ under } H_0.$$

and the critical value of 5% (1%) level of significance is 1.645 (2.33).

(c) To test $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 < \mu_2$ (one-tailed test), the test statistic Z_{Cal} is same as in (b) above, however, the critical value for 5% (or 1%) level of significance is – 1.645 (or – 2.33).

Case II : If σ_1 and σ_2 are not known, their esimates based on samples are used. This category of tests can be further divided into two sub-groups.

Small sample tests (when either n_1 or n_2 or both are less than or equal to 30). For this test $H_0 : \mu_1 = \mu_2$, we use t-test, The respective estimates of σ_1 and σ_2 are given by

$$S_1 = \sqrt{\frac{\sum(X_{1i} - \bar{X}_1)^2}{n_1 - 1}} = S_1 \sqrt{\frac{n_1}{n_1 - 1}} \text{ and } s_2 = \sqrt{\frac{\sum(X_{2i} - \bar{X}_2)^2}{n_2 - 1}} = S_2 \sqrt{\frac{n_2}{n_2 - 1}} .$$

This test is more restrictive because it is based on the assumption that the two samples are drawn from indepedent normal populations with equal standard deviations. i.e. $\sigma_1 = \sigma_2 = \sigma$ (say). The pooled estimate of σ, denoted by s, is defined as

$$s = \sqrt{\frac{\sum(X_{1i} - \bar{X}_1)^2 + \sum(X_{2i} - \bar{X}_2)^2}{n_1 + n_2 - 2}} .$$

$$= \sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

(a) To test $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$ (two-tailed test), the test statistic is

$$t = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}} = \frac{(\bar{X}_1 - \bar{X}_2)}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{(\bar{X}_1 - \bar{X}_2)}{s} \times \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$, which folows the t-distribution with $(n_1 + n_2 - 2)$ d.f.

(b) To test $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 > \mu_2$ (one-tailed test), the test statistic is $t = \frac{(\bar{X}_1 - \bar{X}_2)}{s} \times \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$.

(c) To test $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 < \mu_2$ (one-tailed test), the test statistic, i.e. t is same as in (b). This value is compared with the negative t value.

2. Large Sample test (when each of n_1 and n_2 is geater than 30)

In this case σ_1 and σ_2 are estimated by their repective sample standard deviations S_1 and S_2. The test statistic for two and one-tailed test is

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim N(0, 1)$$ and the remaining procedure is same as above.

Remark :

100 $(1 - \alpha)$% confidence limits for $h_1 - h_2$ are given by $(\bar{X}_1 - \bar{X}_2) \pm Z_{\alpha/2}$ S.E. $_{\bar{X}_1 - \bar{X}_2}$. If the two samples are drawn from populations with same standard deviation, i.e. $\sigma_1 = \sigma_1 = \sigma$ (say), then S.E.$_{\bar{X}_1 - \bar{X}_2} = \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ for problems covered under case I and S.E.$_{\bar{X}_1 - \bar{X}_2} = S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ for problems covered under case II. For large sample tests, σ can also be estimated by S as

$$S = \sqrt{\frac{\Sigma(X_{1i} - \bar{X}_1)^2 + \sum(X_{2i} - \bar{X}_2)^2}{n_1 + n_2}} = \sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2}}.$$

Example 4.

Two types of batteries are tested for their length of life and the following data are obtained.

	No. of sample	Mean life (inches)	Variance
Type - A	9	600	121
Type - B	8	640	144

Is there a significant difference in the two means ? Value of t for 15 d.f. at 5% level is 2.131.

Solution : Let X_1 and X_2 be the life of two types of batteries A and B with mean μ_1 and μ_2 respectively.

Here $H_0 : \mu_1 = \mu_2$, against $H_1 : \mu_1 \neq \mu_2$.

To test the above null hypothesis the appropriate test statistic is

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2} \text{ under the null hypothesis,}$$

where the pooled estimate of σ, denoted by s, is defined as

$$s = \sqrt{\frac{\sum(X_{1i} - \bar{X}_1)^2 + \sum(X_{2i} - \bar{X}_2)^2}{n_1 + n_2 - 2}} = \sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}}$$

In this problem

$n_1 = 9$ $\qquad$ $n_2 = 8$

$\bar{X}_2 = 600$ $\qquad$ $\bar{X}_2 = 640$

$S_1^2 = 121$ $\qquad$ $S_2^2 = 144$

Hence,

$$s = \sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}} = \sqrt{149.4} = 12.2$$

Therefore, $t_{Cal} = \dfrac{600 - 640}{12.2\sqrt{\frac{1}{9} + \frac{1}{8}}} = -6.7$

Since the observed value of | t | (6.7) is greater than the tabulated value of t (2.131) at 5% level of significance so the null hypothesis H_0 is rejected at 5% level of significance. So there is a significant difference between the two means of lives of batteries at 5% level of significance.

8.3.3 Test of Hypothesis Concerning Equality of Two Means (σ_1 and σ_2 unknown for a bivariate population)

Suppose we have a random sample of n pairs of observations from a bivariate normal population with unknown means μ_1, μ_2 and standard deviations σ_1 and σ_2. The population standard deviations are not assumed to be equal. Suppose the paired data are available

(X_i, Y_i), i = 1, 2, 3, ..., n, j = 1, 2, 3..... m_1

Step I. (Hypothesis Formulation)

We set up the null hypothesis and alternative hypothesis on the basis of the above problem. That is, $H_0 : \mu_1 = \mu_2$ against alternatives

a) $H_1 : \mu_1 > \mu_2$

or b) $H_1 : \mu_1 < \mu_2$

or c) $H_1 : \mu_1 \neq \mu_2$

Step II. (Test Statistic)

Let $U_i = X_i - Y_i$, difference in the values of X and Y for the i-th pair, i = 1, 2, 3, ..., n. To test the above null hypothesis we consider the appropriate test statistic $t = \frac{\sqrt{n}\,\bar{U}}{S_U}$ which follows the t-distribution with (n–1) degrees of freedom under the null hypothesis, where $S_U = \sqrt{\frac{\sum(U_i - \bar{U})^2}{n-1}}$.

Step III. (Computation)

Let the value of this statistic calculated from the sample be denoted as $t_{Cal} = \frac{\sqrt{n}\bar{U}}{S_U}$ (say).

Step IV. (Conclusion)

(c) Reject the null hypothesis $H_0 : \mu_1 = \mu_2$ against the alternative $H_1 : \mu_1 > \mu_2$ at α level of significance if $t_{Cal} > t_{\alpha, n-1}$.

(d) Reject the null hypothesis $H_0 : \mu_1 = \mu_2$ against the alternative $H_1 : \mu_1 < \mu_2$ at α level of significance if $t_{Cal} < -t_{\alpha, n-1}$ and

(e) Reject the null hypothesis $H_0 : \mu_1 = \mu_2$ against the alternative $H_1 : \mu_1 \neq \mu_2$ at α level of significance if $| t_{Cal} | > t_{\alpha/2, n-1}$

Here $t_{\alpha ; n-1}$ denotes the upper α - point of a t-distribution with $n - 1$ degrees of freedom. That is, $P(t > t_{\alpha, n-1}) = \alpha$.

Note : The above test is known as Paired t-test which may be viewed as Student's t-test with $n - 1$ degrees of freedom.

8.4 Test of Hypothesis for Proportion

8.4.1 Test of Hypothesis for Specified Population Proportion

Setp I. (Hypothesis Formulation)

We set up the null hypothesis and alternative hypothesis on the basis of the given problem. That is, the null hypothesis to be tested is

$H_0 : \pi_1 = \pi_2$ against alternatives a) $H_1 : \pi > \pi_0$

or b) $H_1 : \pi < \pi_0$

or c) $H_1 : \pi \neq \pi_0$

Step II. (Test Statistic)

To test the above null hypothesis we consider the appropriate test statistic

$$Z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} = (p - \pi_0)\sqrt{\frac{n}{\pi_0(1-\pi_0)}} \sim N(0,1)$$ under the null hypothesis for sufficiently large n.

Step III. (Computation)

Let the value of this statistic calculated from sample be denoted as

$$Z_{Cal} = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \text{ (say)}.$$

Step IV. (Conclusion)

(c) Reject the null hypothesis $H_0 : \pi = \pi_0$ against the alternative $H_1 : \pi > \pi_0$ at α level of significance if $Z_{Cal} > Z_\alpha$.

(d) Reject the null hypothesis $H_0 : \mu = \pi_0$ against the alternative $H_1 : \pi < \pi_0$ at α level of significance if $Z_{Cal} < -Z_\alpha$ and

(e) Reject the null hypothesis $H_0 : \pi = \pi_0$ against the alternative $H_1 : \pi \neq \pi_0$ at α level of significance if $|Z_{Cal}| > Z_{\alpha/2}$.

Here Z_α denotes the upper α - point of the standard normal distribution. That is, $P(Z > Z_\alpha) = \alpha$, where Z is a standard normal variate.

Remark : The $100(1 - \alpha)\%$ confidence limits for p are $\pi \pm Z_{\alpha/2}$S.E.(p).

Example 5 : A cerain controlled process produces 15 percent defective items. A supplier of a basic raw material claims that the use of his material would reduce the fraction of defective. On making a production trial run with the new material, it was found that from an output of 400 units 52 were defective. Would you accept the supplier's claim?

Solution. Let X be the random variable denoting the number of defectives in a sample of size n. We assume that X follows the binomial distribution with parameters n and π.

Here we have to test the null hypothesis

$H_0 : \pi = 0.15$ against the alternative $H_1 : \pi < 0.15$.

$p = X/n = 52 / 400 = 0.13$

$$Z_{Cal} = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} = \frac{0.13 - 0.15}{\sqrt{\frac{0.15 \times 0.85}{400}}} = -1.12 .$$

$Z_{0.05} = 1.645$.

i.e. at 5% level of significance the supplier's claim that the new material will reduce the fraction of defective is rejected at 5% level of significance.

Example 6 : A random sample of 100 items taken from a large batch of articles contains 5% defective items. (a) Set up 96% confidence limit for the proportion of

defective items in a batch, (b) If the batch contains 2, 697 items, set up the 95% confidence limits for the proportion of defective items.

Solution : Here n = 100.

p = proportion of defectives in the sample = 5/100 = 0.05

(a) Here estimate of S.E. of p is given by

$$SE\ (p) = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.05\times.95}{100}} = 0.02179 \cong 0.022.$$

From the table of the normal distribution

$$Z_{0.02} = 2.05.$$

Hence 96% confide nce limits for the population proportion of defectives are :

$$p \pm Z_{0.02}\ S.E.(p) = 0.05 \pm 2.05 \times 0.022 = (0.005, 0.095)$$

(b) If N = 2669, then 95% confidence limits for p are given by

$$p \pm Z_{0.025}\ SE(p) = p \pm 1.96 \times \sqrt{\frac{(N-n)pq}{N(n-1)}}.$$

$$\left[\sqrt{\frac{pq}{n}\frac{(N-n)}{N-1}} \text{ is biased but } \sqrt{\frac{pq}{n-1}\frac{(N-n)}{N}} \text{ is unbiased }\right]$$

$$= 0.050 \pm 1.96 \times \sqrt{\frac{(2669-100)}{2669} \times \frac{0.05\times0.95}{99}}$$

$$= 0.050 \pm 1.96 \times 0.0215 = 0.050 \pm 0.042$$

$$= (0.008, 0.092)$$

8.4.2 Test of Hypothesis Concerning Equality of Proportions

Setp I. (Hypothesis Formulation)

We set up the null hypothesis and alternative hypothesis on the basis of the given problem. That is, the null hypothesis to be tested is

$H_0 : \pi_1 = \pi_2$ against the alternatives a) $H_1 : \pi_1 > \pi_2$

or b) $H_1 : \pi_1 < \pi_2$

or c) $H_1 : \mu_1 \neq \pi_2$

Step II. (Test Statistic)

To test the above null hypothesis we consider the appropriate test statistic

$$Z = \frac{(p_1 - p_2)}{\sqrt{\pi(1-\pi)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = (p_1 - p_2)\sqrt{\frac{n_1 n_2}{\pi(1-\pi)(n_1+n_2)}} \sim N(0, 1)$$ under the null

hypothesis for sufficiently large n_1 and n_2 under the assumption that $\pi_1 = \pi_2 = \pi$. where π is known. Often population proportion π is unknown and it is estimated on the basis of samples. The pooled estimate of π, denoted by p, is given by $p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$.

Thus, the test statistic becomes $Z = (p_1 - p_2)\sqrt{\frac{n_1 n_2}{p(1-p)(n_1+n_2)}}$.

Setp III. (Computation)

Let the value of this statistic calculated from the sample be denoted as

$$Z_{Cal} = (p_1 - p_2)\sqrt{\frac{n_1 n_2}{p(1-p)(n_1+n_2)}} \text{ (say)}$$

Step IV. (Conclusion)

(c) Reject the null hypothesis $H_0 : \pi_1 = \pi_2$ against the alternative $H_1 : \pi_1 > \pi_2$ at α level of significance if $Z_{Cal} > Z_\alpha$.

(d) Reject the null hypothesis $H_0 : \pi_1 = \pi_2$ against the alternative $H_1 : \pi_1 < \pi_2$ at α level of significance if $Z_{Cal} < -Z_\alpha$ and

(e) Reject the null hypothesis $H_0 : \pi_1 = \pi_2$ against the alternative $H_1 : \pi_1 \neq \pi_2$ at α level of significance if $|Z_{Cal}| > Z_\alpha$.

Here Z_α denotes the upper α - point of standard normal distribution. That is, $P(Z > Z_\alpha) = \alpha$ where Z is a standard normal variate.

Remark : The $100(1-\alpha)\%$ confidence limits for $\pi_1 - \pi_2$ are $(p_1 - p_2) \pm Z_{\alpha/2}$ S.E.$(p_1 - p_2)$.

Example 7 : A survey of television audiance in a big city revealed that a particular programme was liked by 50 out of 200 males and 80 out of 250 females. Test the hypothesis that whether there is a real difference of opinion about the programme between males and females.

Solution : Let π_1 and π_2 be the proportion of males and females who liked the particular television prgaramme. The null hypothesis to be tested is

$H_0 : \pi_1 = \pi_2$ against alternative $H_1 : \pi_1 \neq \pi_2$.

To test the above null hypothesis we condsider the appropriate test statistic

$$Z = \frac{(p_1 - p_2)}{\sqrt{\pi(1-\pi)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = (p_1 - p_2)\sqrt{\frac{n_1 n_2}{\pi(1-\pi)(n_1 + n_2)}} \sim N(0,1) \text{ under the null}$$

hypothesis for large n_1 and n_2 under the assumption that $\pi_1 = \pi_2 = \pi$ where π is known. Often population proportion π is unknown and it si estimated on the basis of samples.

The pooled estimate of π denoted by p is given by $p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$.

Here $p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{50 + 80}{200 + 250} = \frac{13}{45}$

So, $Z_{Cal} = \frac{\frac{80}{250} - \frac{50}{200}}{\sqrt{pq\left(\frac{1}{250} + \frac{1}{200}\right)}} = \frac{0.32 - 0.25}{0.04305} = 0.958$.

Since the computed value of Z (0.958) is less than the tabulated value of Z (1.96) at 5% level of isgnificance, so the null hypothesis is accepted. That is, there is no real difference of opinion about the programme between males and females.

Example 8 : Obtain the 95% confidence limits for the proportion of success in a binomial population.

Solution : Let the parameter π denote the proportion of successes in population. Further, p denote the proportion of successes in n ($\neq 50$) trials. We know that the sampling distribution of p will be approimately normal with mean p and standard error $\sqrt{\frac{\pi(1-\pi)}{n}}$

Since π is not known, therefore, its estimator p is used in the estimation of standard error of p, i.e. $S.E.(p) = \sqrt{\frac{p(1-p)}{n}}$.

Thus, the 95% confidence interval for p is given by

$$P\left(p-1.96\sqrt{\frac{p(1-p)}{n}} \leq \pi \leq p+1.96\sqrt{\frac{p(1-p)}{n}}\right)=0.95.$$

This gives the 95% confidence limits as $p \pm 1.96\sqrt{\frac{p(1-p)}{n}}$.

Example 9 : In a newspaper article of 1600 words in Hindi, 64% of the words were found to be of Sanskrit origin. Assuming that the simple sampling conditions hold good, estimate the confidence limits for the proportion of Sanskrit words in the writer's vocabulary.

Solution : Let π be the proportion of Sanskrit words in the writer's vocabulary. Corresponding proportion in the sample is given as p = 0.64

$$\therefore S.E.(p)=\sqrt{\frac{0.64\times0.36}{1600}}=\frac{0.48}{40}=0.012.$$

We known that almost whole of the distribution lies between 3σ limits. Therefore, the confidence interval is given by

$$P[p - 3S.E.(p) \leq \pi \leq p + 3S.E.(p)] = 0.9973$$

Thus, the 99% confidence limits are 0.609 (0.64 – 2.58 × 0.012) and 0.671 (=0.64 + 2.58 × 0.012) respectively.

Hence, the proportion of Sanskrit words in the writer's vocabulary are between 60.9% and 67.1%.

8.5 Test of Hypothesis Concerning Population Standard Deviation

Step I. (Hypothesis Formulation)

We set up the null hypothesis and alternative hypothesis on the basis of the given problem. That is, the null hpothesis to be tested is

$H_0 : \sigma = \sigma_0$ against the alternatives a) $H_1 : \sigma_1 > \sigma_0$

or b) $H_1 : \sigma < \sigma_0$

or c) $H_1 : \sigma \neq \sigma_0$.

Step II. (Test Statistic)

To test the above null hypothesis we consider the appropriate test statistic

$\chi^2 = \frac{\Sigma(X_i - \bar{X})^2}{\sigma_0^2} = \frac{nS^2}{\sigma_0^2}$ follows a χ^2 - variate with $(n - 1)$ degrees of freedom under the null hypothesis.

Setp III. (Computation)

Let the value of this statistic calculated from the sample be denoted as

$\chi^2_{Cal} = \frac{\sum(X_i - \bar{X})^2}{\sigma_0^2} = \frac{nS^2}{\sigma_0^2}$ (say).

Step IV. (Conclusion)

(c) Reject the null hypothesis $H_0 : \sigma = \sigma_0$ against the alternative $H_1 : \sigma > \sigma_0$ at α level of significance if $\chi^2_{Cal} > \chi^2_{\alpha;(n-1)}$.

(d) Reject the null hypothesis $H_0 : \sigma = \sigma_0$ against the alternative $H_1 : \sigma < \sigma_0$ at α level of significance if $\chi^2_{Cal} > \chi^2_{1-\alpha;(n-1)}$ and

(e) Reject the null hypothesis $H_0 : \sigma = \sigma_0$ against the alternative $H_1 : \sigma \neq \sigma_0$ at α level of significance if $\chi^2_{Cal} < \chi^2_{\frac{\alpha}{2};(n-1)}$ or $\chi^2_{Cal} > \chi^2_{1-\frac{\alpha}{2};(n-1)}$.

Here $\chi^2_{\alpha;\,(n-1)}$ denotes the uper α - point of a χ^2 - variate with $(n-1)$ degrees of fredom. That is, $P\left(\chi^2 > \chi^2_{\alpha;(n-1)}\right) = \alpha$.

Note : It can be shown that for large sample ($n > 30$), the sampling distribution of S is approbiatelity normal with mean σ and standard error $\frac{\sigma}{\sqrt{2n}}$. Thus

$Z = \frac{(S-\sigma)\sqrt{2n}}{\sigma} \sim N(0, 1)$ for sufficiently large value of n.

Alternativcly, using Fisher's approximation, we can say that when $n > 30$ the statistic $\sqrt{2\chi^2}$ follows a normal distribution with mean $\sqrt{2n-1}$ and standard error unity. Thus $Z = \sqrt{2\chi^2} - \sqrt{2n}$ can be taken as a standard normal variate for sufficiently alrge values of n.

8.6 Test of Hypothesis Concerning the Equality of Standard Deviations

Step I. (Hypothesis Formulation)

We set up the null hypothesis and the alternativc hypothesis on the basis of the given problem. That is, the null hypothesis to be tested is

$H_0 : \sigma_1 = \sigma_2$ against alternative $H_1 : \sigma_1 > \sigma_2$.

Step II. (Test Statistic)

To test the above null hypothesis we consider thc appropriate test statistic $F = \frac{s_1^2 / \sigma_1^2}{s_2^2 / \sigma_2^2}$ which would becomc $\frac{s_1^2}{s_2^2}$ and under H_0 it follows the F – distribution with v_1 $(= n_1 - 1)$ and v_2 $(= n_2 - 1)$ degrees of freedom.

Setp III. (Computation)

Lct the value of this statistic calculated from the sample be denoted as $F_{Cal} = \frac{s_1^2}{S_2^2}$ (say).

Setp IV. (Conclusion)

Reject the null hypothesis $H_0 : \sigma_1 = \sigma_2$ against the alternative $H_1 : \sigma_1 > \sigma_2$ at α level of significance if $F_{Cal} > F_{\alpha;(n_1-1)(n_2-1)}$.

Here $F_{\alpha:(n_1-1)(n_2-1)}$ denotes the upper α - point of an F – variate with $(n_1 - 1)$ and $(n_2 - 1)$ degrees of freedom. That is, $P\left(F > F_{\alpha;(n_1-1)(n_2-1)}\right) = \alpha$.

Remarks :

1. We can write $S_1^2 = \frac{1}{n_1 - 1}\sum(X_{1i} - \bar{X}_1)^2 = \frac{n_1}{n_1 - 1}S_1^2 = \frac{1}{n_1 - 1}\left(\sum X_{1i}^2 - \frac{(\sum X_{1i})^2}{n_1}\right)$

   and $s_2^2 = \frac{1}{n_2 - 1}\sum(X_{2i} - \bar{X}_1)^2 = \frac{n_2}{n_2 - 1}S_2^2 = \frac{1}{n_2 - 1}\left(\sum X_{2i}^2 - \frac{(\sum X_{2i})^2}{n_2}\right)$.

2. In the variance ratio $F = \frac{s_1^2}{s_2^2}$, we take, by convenion, the largest of the two sample variances as s_1^2. Thus, this test is always a one-tailed test with critical region at the right hand tail of the F - distribution.

3. The 100 $(1 - \alpha)$% confidence limits for the variance ratio $\frac{\sigma_1^2}{\sigma_2^2}$ are given by

$$P\left(\frac{s_1^2}{s_2^2} \cdot \frac{1}{F_{\alpha/2}} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{s_1^2}{s_2^2} \cdot \frac{1}{F_{1-\alpha/2}}\right) = 1 - \alpha .$$

8.7 Frequeny Chi-square (Pearsonian χ^2)

8.7.1 Test for Goodness of Fit

The use of chi-square (χ^2) test was first devised by Karl Pearson to decide whether the observations are in good agreement with a hypothetical distribution i.e., whether the sample may be supposed to have come from a specified population. The observed

values (f_0) for different classes are compared with expected values (f_e) forming the test statistic.

$$\therefore \chi^2 = \Sigma \frac{(f_0 - f_e)^2}{f_e} \sim \chi^2_{k-1}.$$

This is called the goodness of fit Chi-square or Pearsonian Chi-square or freqhency Chi-square distribution with k-1 degrees of freedom under the null hypothesis where k is the number of classes. We reject the null hypothesis if the observed value of the statistic exceeds the tabulated value of χ^2 at a particular level.

Example 9 : A die was thrown 60 times with the following results.

Face	1	2	3	4	5	6	Total
Frequency	6	10	8	13	11	12	60

Are the data consistent with the hypothesis that the die is unbiased ? (Gven χ^2 = 15.09 for 5 degrees of freedom at 0.01 level)

Solution : Let us consider that the null hypothesis is that the die is unbiased. Then the probability of each face is 1/6 and the expected frequency is 10 for each.

Values	Observed frequency (f_0)	Expected frequency (f_e)	$(f_0 - f_e)^2/f_0$
1	6	10	1.6
2	10	10	0
3	8	10	0.4
4	13	10	0.9
5	11	10	0.1
6	12	10	0.4
Total	**60**	**60**	**3.4**

$$\therefore \chi^2 = \sum \frac{(f_0 - f_e)^2}{f_e} = 3.4.$$

There are 6 classes and hence degree of freedom = 6 – 1 = 5.

Since the observed value of χ^2 is less than the tabulated value of $\chi^2_{0.01} = 15.09$,

we accept the null hypothesis at 1% level of significance and conclude that the die is unbiased.

Example 10. 5 identical coins are tossed 320 times and the number of heads appearing each time is recorded. The results are :

No. of heads	:	0	1	2	3	4	5	Total
Frequency	:	14	45	80	112	61	8	320

would you decide that the coins are biased?

(Given $\chi^2_{0.05} = 11.07$, $\chi^2_{0.01} = 15.09$ for 5 degrees of freedom)

Solution. Let X follow the binomial distribution with parameters n and p.

Here the null hypotheis $H_0 : p = 1/2$ against $H_1 : p \neq 1/2$.

That is, under the null hypothesis the p.m.f. is $p(x) = {}^5C_x \left(\frac{1}{2}\right)^5$.

No. of heads (x)	f_0	p(x)	f_e(Np(x)	$[f_0 - f_e]^2 / f_e$
0	14	0.03125	10	1.60
1	45	0.15625	50	0.50
2	80	0.31250	100	4.00
3	112	0.31250	100	1.44
4	61	0.15625	50	2.42
5	8	0.03125	10	0.40
Total	**N=320**	**1**	**320**	**10.36**

Since the observed value (10.36) is less than the tabulated value (11.07) at 5% level of significance, so the H_0 is accepted at 5% level of significance. Hence, we can decide that the coins can be treated as unbiased.

8.7.1 Test for Independence of Attributes

When observations are classified according to two attributes and arranged in a two-way table, the display is put in terms of a contingeny table.

Two-Way Contingency Table

Attribute B	Attribute A	Total
	A_1 A_2 ..... A_n	
B_1	O_{11} O_{12} ..... O_{1n}	R_1
B_2	O_{21} O_{22}..... O_{2n}	R_1
.	.	.
.	.	.
B_m	O_{m1} O_{m2} ..... O_{mn}	R_m
Total	C_1 C_2 ..... C_n	N

Here it may be noted that the attributes A and B have been classified into mutually exclusive categories. The value O_{ij} represents the frequency of the observation corresponding to the i-th row and j-th column. The expected frequency E_{ij} is given by $E_{ij} = \frac{R_i}{N} \times \frac{C_j}{N} \times N = \frac{R_i \times C_j}{N}$, i = 1, 2, 3, ..., n .

Here the null hypothesis H_0 : A and B are independent
against the alternative H_1 : A and B are dependent
The test statistic under the null hypothesis for the test is

$$\chi^2 = \sum \frac{(O-E)^2}{E} \sim \chi^2_{(m-1)(n-1)} .$$

We reject the null hypothesis at α% level of significance if observed $\chi^2 > \chi^2_{\alpha,\,(m-1)(n-1)}$.

Example 11 : In a recent diet survey, the following results were obtained in an Indian city :

No. of families	Hindus	Muslims	Total
Tea takers	1236	164	1400
Non-tea takers	564	36	600
Total	**1800**	**200**	**2000**

Discuss whether there is any sigificant difference between the two communities in the matter of taking tea. Use 5% level of significance.

Solution : The null hypothesis that is to be tested can be written as H_0 : There is no difference between the two communities in the matter of taking tea.

Using the direct formula, we have $\chi^2 = \frac{2000(1236\times36 - 164\times564)^2}{1400\times1800\times200\times600} = 15.24$.

The value of $\chi^2$from the table for 1 d.f. and at 5% level of significance is 3.84. Since the calculated value is greater than the tabulated value, H_0 is rejected. That is, there is a significant difference between the two communities in the matter of taking tea at 5% level of significance.

Example 12 : A certain drug is claimed to be effective in curing clods. In an expenriment on 500 persons with clods, half of them were given the drug. The patients' reaction to the treatment are recorded in the following table.

Treatment	**Helped**	**Reaction**	**No efect**	**Total**
Drug	150	30	70	250
Sugar Pills	130	40	80	250
Total	**280**	**70**	**150**	**500**

On the basis of the data, can it be concluded that there is a significant difference in the effect of the drug and sugar pills? (Given $\chi^2_{0.05,\,2} = 5.99$).

Solution : Let us take the null hypothesis that there is no significant difference in the effect of the drug and sugar pills.

The contingency table is of size 2 × 3, the degree of freedom would be (2 – 1) (3 – 1) = 2. The expected frequencies can be calculated in the follwoing way :

$$E_{11} = \frac{280\times250}{500} = 140,\ E_{11} = \frac{280\times250}{500} = 140, \text{ and so on.}$$

Contingency table for expected frequencies is as follows :

Treatment	**Helped**	**Reaction**	**No efect**	**Total**
Drug	140	35	75	250
Sugar Pills	140	35	75	250
Total	**280**	**70**	**150**	**500**

Arranging the observed and the expected frequencies in the following table we calculate the value of χ^2 test statistic.

Cell (i,j)	Observed frequency (O)	Expected frequency (E)	$(O - E)^2/E$
1,1	150	140	0.714
2,1	130	140	0.714
1,2	30	35	0.714
2,2	40	35	0.714
1,3	70	75	0.333
1,3	80	75	0.333
Total	**500**	**500**	**3.522**

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 3.522 .$$

Since, the observed $\chi^2 < \chi^2_{0.05,2}$, therefore, we accept the null hypothesis at 5% level and conclude that there is no significant difference in the effect of the drug and suggar pills.

8.8 Exercises

1. What is test of significance ? Explain the procedure generally followed in testing of hypothesis.
2. Distinguish between (a) critical region and acceptance regions, (b) null hypothesis and alternative hypothesis, (c) one-tailed and two-tailed test, (d) type I error and type II error.
3. Explain clearly the procedure of testing hypothesis. Also point out the assumptions in hypothesis testing in large samples.
4. How does small sampling theory differ from large sampling theory?
5. Explain the following terms : test statistic, level of significance, confidence level and power of a test.

6. Give some important applications of 't' test and explain how it helps in business decision making.

7. Discuss the F-test for testing the equality of two variances.

8. What is chi-square test? Explain its important uses with the help of examples.

9. A sample of 25 male students is found to have a mean height of 171.38cm. Can it be reasonably regarded as a sample from a population with a mean height of 171.17 cms and a standard deviation of 3.30 cms?

10. A toothpase company conducted a survey and found that it could sell only 60 tubes on an average per month per shop. Immediately, the company advertised heavily in several media and after 3 months again conducted a survey and found that the mean sales was 83 tubes with a standard deviation of 10 tubes in a sample of 20 shops. Can it be concluded that the advertiesement is effective? ($t_{0.05;\ 19} = 1.729$, $t_{0.01\ ;\ 19} = 2.861$)

11. In a survey at a super market, the following number of people were observed purchasing different brands of coffee :

A	B	C	D	E
74	53	81	70	82

Do these data support the hypothesis that the population of coffee buyers prefer each of the five brands equally?

Hint : $\chi^2_4 = \dfrac{2^2+19^2+9^2+2^2+10^2}{72} = 7.64$

12. A sample of 540 households was selected to study the occupational pattern of the father and the son. The number of households obtained has been tabulated below. Test the hypothesis that the son's occupation is independent of the father's occupation.

		Son's occupation			
		F	B	S	M
Father's occupation	Farming	24	97	62	58
	Business	22	28	30	41
	Services	32	10	11	20
	Miscellancous	38	25	14	28

($\chi^2_{9;\ 0.05} = 16.919$, $\chi^2_{9;\ 0.01} = 21.666$).

13. An automobile manufacturing firm is bringing out a new model. In order to map out its advertising campaign, it wants to determine whether the model will appeal most to a particular age group or equally to all age groups. The firm conducted a survey and the results are summarised below :

	Age group			
	Below 20	20-39	40-59	60 and above
Liked	146	78	48	28
Disliked	54	52	32	62

What conclusion would you draw from the above data? ($\chi^2_{3\,;\,0.05} = 7.815$, $\chi^2_{3;\,0.01} = 11.341$)

14. A man buys 15 electric bulbs of 'Philips' make and another 10 of the 'GE' make. He finds that the Philips bulbs give an average life of 1200 hours with S.D. of 60 hours and the GE bulbs give an average of 1242 hours with an S.D. of 80 hours. Is there a significant difference between the two makes?

($t_{23;\,0.025} = 2.069$, $t_{23\,;\,0.005} = 2.806$)

15. The sales data of an item in six shops before and after a special promotional campaign are as follows :

Shops :	A	B	C	D	E	F
Before campaign :	53	28	31	48	50	42
After campaign :	58	29	30	55	56	45

Can the campaign be judged to be a success ? ($t_{5\,;\,0.05} = 2.015$)

16. Two types of scooters manufactured in India are tested for petrol mileage. One group consisting of 12 scooters with average mileage of 44 km/lt and of standard deviation of 2 km while the other group consisting of 10 scooters with an average mileage of 50 km/ lt and standard deviation of 1.5 km of petrol. Test whether statistically there exists a significant difference in the petrol consumption of two types of scooters. ($t_{20\,;\,0.025} = 2.086$, $t_{20\,;\,0.005} = 2.845$)

17. Two laboratories A and B carry out independent estimates of fat content in ice-cream made by a firm. A sample is taken from each population, halved and the

separate halves sent to the two laboratories. The fat content obtained by the laboratories is recorded below :

Batch No :	1	2	3	4	5	6	7	8	9	10
Lab A :	3	5	7	3	8	6	9	3	7	8
LabB :	9	8	8	4	7	7	9	6	6	6

Is there a significant difference between the mean fat content obtained by the two laboratories A and B? ($t_{9\,;\,0.025} = 2.262$, $t_{9\,;\,0.005} = 3.250$)

18. A company making a brand of detergent and toilet soap wanted to compare the expenses incurred on sales promotion for these two products. The data for the preceding year were retrieved from the books of accounts of these two products and they are reproduced below :

Expenditure in Rs. Thousand

Months Product	1	2	3	4	5	6	7	8	9	10	11	12
Toilet soap	55	80	50	60	50	60	70	45	50	60	60	70
Detergent	50	25	70	45	60	55	45	60	55	55	45	35

Further, suppose that both the products had offered equal amount of profitability and turnover. Then verify whether the above sales promotion expenditure are justifiable or not. ($t_{0.025\,;\,11} = 2.20$, $t_{0.005\,;\,11} = 3.11$).

19. The following data were obtained from a test in a laboratory.

Method	sample size	sample variance
A	10	1296
B	15	784

Test whether there is any difference in the variances of two methods at 5% level.

($F_{9\,,\,14;\,0.05} = 2.65$, $F_{9,\,14\,;\,0.01} = 4.03$)

20. A random sample of 15 observations gave an unbiased estimator $s^2 = 12.63$ of the population variance σ^2. May the sample be reasonably regarded as drawn from a normal population with variance 8 ? Test at 5% level of significance.

($\chi^2_{0.05;\,14} = 23.68$, $\chi^2_{0.01;14} = 29.14$)

21. A stock broker claims that he can predict with 80% accuracy whether the values of a stock will rise or fall during the coming month. As a test he predicts the outcome of 40 stocks and is correct in 28 of the predictions. Does the evidence support the stock broker's claim?

22. 500 units from a factory are inspected and 12 are found to be defective. Similarly, 800 units from another factory are inspected and 17 are found to be defective. Can it be concluded that production in the second factory is better than in the first?

23. Determine the sample size for estimating the true weight of tea containers from (i) a large number of containers and (ii) from 1000 containers so that the estimate should be within 10 gms of the true average weight. Variance is 40 gms (on the basis of past record).

Hint :

$E = |\bar{X} - \mu| = 10$ $\quad\quad$ $\sigma^2 = 40$

$n = \dfrac{\sigma^2 z_{\alpha/2}^2}{E^2}$ $\quad\quad$ for SRSWR

$n = \dfrac{N(\sigma^2 z_{\alpha/2}^2)}{NE^2 + \sigma^2 z_{\alpha/2}^2}$ $\quad\quad$ for SRSWR

For proportion

$n = \dfrac{p(1-p) z_{\alpha/2}^2}{E^2}$ $\quad\quad$ for SRSWR

$n = \dfrac{Np(1-p) z_{\alpha/2}^2}{NE^2 + p(1-p) z_{\alpha/2}^2}$ $\quad\quad$ for SRSWR

References

Aczel, A. D. and Sounderpandian, J. (2002). *Complete Business Statistics,* Tata McGraw Hill, New Delhi.

Agarwal, B. L. (1996). *Basic Statistics,* New Age International, New Delhi.

Agarwal, D. R. (2003). *Quantitative Methods,* Vrinda Publications, New Delhi.

Agarwal, B. M. (2003). *Business Statistics,* Sultan Chand and Sons, New Delhi,

Anderson, D. R., Sweeney D. J. and Williams, T. A (2002). *Statistics for Business and Economics,* Thomson Asia Pvt Ltd, New Delhi.

Beri, G. C. (2005). *Statistics for Management,* Tata McGraw Hill, New Delhi.

Bhardwaj, R. S. (1999). *Business Statistics* .Excel Books, New Delhi.

Chandan, J. S., Singh, J. and. Khanna, K. K. (1995). *Business Statistics,* Vikash Publishing House Pvt.Ltd, New Delhi.

Das, N. G. (2002). *Statistical Methods in Commerce, Accountancy & Economics,* M. Das & Co., Kolkata.

Fleming, M. C. and. Nellis, J. C. (2002). *The Essence of Statistics for Business,* Prentice-Hall of India, New Delhi.

Giri, P. K. and Bannerjee, J. (2002). *Statistical Tools and Techniques,* Academic Publishers, Kolkata.

Gupta, S. P. (2003). *Statistical Methods* , Sultan Chand and Sons, New Delhi.

Gupta, S. P. and Gupta, M. P. (1999). *Business Statistics,* Sultan Chand and Sons, New Delhi.

Gun, A. M., Gupta, M. K. and Dasgupta, B. (2002). *Fundamentals of Statistics,* Vol. 1, The World Press, Kolkata.

Gupta, S. C. (2002). *Fundamentals of Statistics,* Himalaya Publishing House, New Delhi.

Hooda, R. P. (2003). *Statistics for Business and Economics,* Macmillan India Limited, New Delhi

Johnson, R. A. and Wichern, D. W. (2003). *Busines Statistics : Decision Making With Data.,* John Wiley and Sons, New Delhi.

Levin, R. I. and Rubin, D.S.(2001). *Statistics for Management,* Prentice-Hall of India, New Delhi.

Mathai, A. M. and Rathie P. N. (1998). *Probability and Statistics,* Macmillan India Limited, New Delhi

Miller, I and Miller, M. (1999). *Mathematical Statistics,* Prentice-Hall of India, New Delhi.

Mukhopadhyay, P. (2000). *Theory and Methods of Survey Sampling,* Prentice Hall of India, New Delhi.

Pal, N. and Sarkar, S. (2005). *Statistics: Concepts and Tools,* Prentice Hall of India, New Delhi.

Chaudhary, F. S. (1999). *Sample Survey Decisions,* New Age International, New Delhi.

Srivastava, U. K., Shenoy, G.V. and Sharma, S. C. (1996). *Quantitative Techniques for Managerial Decisions,* New Age International, New Delhi.

Shenoy, G. V. and Pant, M. (1994). *Statistical Methods in Business and Social Science,* Macmillan India Limited, New Delhi.

Spiegel, M. R. and Stephens, L. J. (2003). Theory and Problems of Statistics (Schaum's Outline Series), Mc Graw Hill International, New Delhi.

Sheldon, M. R. (2006). Introductory Statistics, Elsevier (Academic Press), Delhi.

Viswanathan, P. K. (2003). Business Statistics: An Application Orientation, Pearson Education, Singapore.

Appendix Table 1 : Area Under standard Normal Curve

(The given proportions indicate area above the given value of Z)

Normal Deviate Z	.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.2	0.4207	0.4168	0.4129	0.4090	0.4062	0.4013	0.3974	0.3936	0.3897	0.3859
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.7	0.2420	0.2389	0.2258	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1822	0.1894	0.1867
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
2.1	0.0179	0.0174	0.0170	0.0156	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
2.8	0.0026	0.0025	0.0024	0.0033	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010

Appendix Table 2: Percentile Values of the Student's t- distribution

df \ 1− α	0.75	0.9	0.95	0.975	0.99	0.995	0.9995
1	1	3.078	6.314	12.706	31.821	63.657	636.619
2	0.816	1.886	2.92	4.303	6.965	9.925	31.598
3	0.765	1.638	2.353	3.182	4.541	5.841	12.941
4	0.741	1.533	2.132	2.776	3.747	4.604	8.61
5	0.727	1.474	2.015	2.571	3.365	4.032	6.859
6	0.718	1.44	1.913	2.447	3.143	3.707	5.959
7	0.711	1.415	1.895	2.365	2.998	3.499	5.405
8	0.706	1.387	1.86	2.306	2.896	3.355	5.041
9	0.703	1.383	1.833	2.262	2.821	3.250	4.781
10	0.700	1.372	1.812	2.228	2.764	3.169	4.587
11	0.697	1.363	1.796	2.201	2.718	3.106	4.437
12	0.695	1.356	1.782	2.179	2.671	3.055	4.318
13	0.694	1.35	1.771	2.160	2.65	3.012	4.221
14	0.692	1.345	1.761	2.145	2.624	2.977	4.140
15	0.691	1.341	1.753	2.131	2.602	2.947	4.073
16	0.690	1.337	1.746	2.12	2.583	2.921	4.015
17	0.689	1.333	1.740	2.11	2.567	2.898	3.965
18	0.688	1.330	1.734	2.101	2.552	2.878	3.992
19	0.688	1.328	1.729	2.093	2.539	2.861	3.883
20	0.687	1.325	1.725	2.086	2.528	2.845	3.850
21	0.686	1.323	1.721	2.08	2.518	2.831	3.819
22	0.686	1.321	1.717	2.074	2.508	2.819	3.792
23	0.685	1.319	1.714	2.069	2.500	2.807	3.767
24	0.685	1.318	1.711	2.064	2.492	2.797	3.745
25	0.684	1.316	1.708	2.060	2.485	2.787	3.725
26	0.684	1.315	1.706	2.056	2.479	2.779	3.707
27	0.684	1.314	1.703	2.052	2.473	2.771	3.690
28	0.683	1.313	1.701	2.048	2.467	2.763	3.674
29	0.683	1.311	1.699	2.045	2.462	2.756	3.659
30	0.683	1.310	1.697	2.042	2.457	2.750	3.646
40	0.681	1.303	1.684	2.021	2.423	2.704	3.551
60	0.699	1.296	1.671	2.000	2.390	2.660	3.460
120	0.677	1.289	1.658	1.980	2.358	2.617	3.373
∞	0.674	1.282	1.645	1.960	2.326	2.576	3.291

Appendix Table 3 : Percentile Values of the the Chi-square distribution

df \ α	0.99	0.98	0.95	0.9	0.8	0.7	0.5	0.05	0.01	0.001
1	...	...	0.00393	0.0158	0.642	0.148	0.455	3.841	6.635	10.827
2	0.0201	0.0401	0.103	0.211	0.446	0.713	2.386	5.991	9.21	13.815
3	0.115	0.185	0.352	0.584	1.005	1.414	2.366	7.815	11.345	16.266
4	0.297	0.429	0.711	1.064	1.649	2.195	3.357	9.488	13.277	18.467
5	0.554	0.752	1.145	1.710	2.343	3.000	4.351	11.07	15.086	20.515
6	0.862	1.134	1.635	2.204	3.07	3.828	5.348	12.592	16.812	22.457
7	1.139	1.564	2.167	2.833	3.822	4.671	6.346	14.067	18.475	24.322
8	1.646	2.032	2.733	3.49	4.594	5.527	7.344	15.507	20.09	26.125
9	2.088	2.532	3.325	4.168	5.38	6.393	8.343	16.919	21.666	27.877
10	2.558	3.059	3.94	4.865	6.179	7.267	9.342	18.307	23.209	29.588
11	3.053	3.609	4.575	5.578	6.989	8.148	10.341	19.675	24.725	31.264
12	3.571	4.178	5.226	6.304	7.807	9.024	11.34	21.026	26.217	32.909
13	4.107	4.765	5.892	7.042	8.634	9.926	12.34	22.362	27.688	34.528
14	4.66	5.368	6.571	7.79	9.467	10.821	13.339	23.685	29.141	36.123
15	5.229	5.985	7.261	8.547	10.307	11.721	14.339	24.996	30.578	37.697
16	5.812	6.614	7.962	9.312	11.152	12.624	15.338	26.296	32.001	39.252
17	6.408	7.255	8.672	10.085	12.002	13.531	16.338	27.587	33.409	40.791
18	7.015	7.906	9.390	10.865	12.857	14.44	17.338	28.869	34.805	42.312
19	7.633	8.567	10.117	11.651	13.716	15.352	18.338	30.144	36.191	43.821
20	8.26	9.237	10.851	12.443	14.578	16.266	19.337	31.41	36.566	45.315
21	8.897	9.915	11.591	13.24	15.445	17.182	20.337	32.641	38.932	46.798
22	9.542	10.6	12.238	14.041	16.314	18.101	21.337	33.924	40.289	48.268
23	10.196	11.293	13.091	14.848	17.187	19.201	22.337	35.172	41.638	49.728
24	10.856	11.992	13.848	15.659	18.062	19.943	23.337	36415	42.98	51.179
25	11.524	12.697	14.611	16.473	18.94	20.867	24.337	37.652	44.314	52.620
26	12.198	13.409	15.379	17.292	19.82	21.792	25.336	38.885	45.642	54.052
27	12.879	14.125	16.151	18.114	20.703	22.719	26.336	40.113	46.963	55.467
28	13.565	14.847	16.928	18.939	21.588	23.647	27.336	41.337	48.278	56.893
29	14.256	15.574	17.708	19.768	22.475	24.577	28.336	42.557	49.588	58.303
30	14.953	16.306	18.493	20.589	23.364	25.508	29.336	43.773	50.892	59.703
40	22.164	23.838	26.509	29.051	32.345	34.872	39.335	55.759	63.692	73.402
50	29.707	31.644	34.764	37.689	41.449	44.313	34.335	67.595	76.154	86.661
60	37.485	39.699	43.188	46.459	50.641	53.809	59.335	79.082	88.379	99.607

Appendix Table 4: 5% percent values of the F-distribution

df(n_2) \ df(n_1)	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	.....
1	161.4	199.5	215.7	224.6	230.2	234	236.8	283.9	240.5	241.9	243.9	245.9	248	249.1	250.1	251.1	252.2	253.3	254.3
2	18.51	19	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.4	19.41	19.43	19.45	19.45	1946	19.47	19.48	19.49	19.5
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.7	8.66	8.64	8.62	8.59	8.57	8.55	8.5
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.8	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.5	4.46	4.43	4.4	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.1	4.06	4	3.94	3.87	3.84	3.81	3.77	3.74	3.7	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.3	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.5	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.9	2.86	2.83	2.79	2.75	2.71
10	4.96	4.1	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.7	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.2	3.09	3.01	2.95	2.9	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.4
12	4.75	3.89	3.49	3.26	3.11	3	2.91	2.85	2.8	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.3
13	4.64	3.81	3.46	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.6	2.53	2.46	2.42	2.38	2.34	2.3	2.25	2.21
14	4.6	3.74	3.34	3.11	2.96	2.85	2.76	2.7	2.65	2.6	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.9	2.79	2.71	2.64	2.59	2.54	2.48	2.4	2.33	2.29	2.25	2.2	2.16	2.11	2.07
16	4.49	3.63	3.34	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.2	2.96	2.81	2.7	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.1	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.9	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.1	2.87	2.71	2.6	2.51	2.45	2.39	2.35	2.28	2.2	2.12	2.08	2.04	1.99	1.95	1.9	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.1	2.05	2.01	1.96	1.92	1.87	1.81
22	4.3	3.44	3.05	2.82	2.66	2.55	2.46	2.4	2.34	2.3	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.8	2.64	2.53	2.44	2.37	2.32	2.27	2.2	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.4	3.01	2.78	2.62	2.51	2.42	2.36	2.3	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.6	2.49	2.4	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.9	1.85	1.8	1.75	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.2	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	4.2	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	4.18	3.33	2.93	2.7	2.55	2.43	2.35	2.28	2.22	2.18	2.1	2.03	1.94	1.9	1.85	1.81	1.75	1.7	1.64
30	4.17	3.32	2.92	2.69	2.45	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.37	2.34	2.25	2.18	2.12	2.08	2	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.01	3.15	2.76	2.53	2.29	2.25	2.17	2.1	2.04	1.99	1.92	1.84	1.75	1.7	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.43	2.21	2.17	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.5	1.43	1.35	1.25
	3.84	3	2.6	2.37	2.19	2.1	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	∞

Appendix Table 5: 1% percent values of the F-distribution

df (n_2)	df (n_1) 1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	....
1	4,052	5,000	5,403	5,625	5,764	5,859	5,928	5,982	6,022	6,056	6,106	6,157	6,209	6,235	6,261	6,287	6,313	6,339	6,366
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.17	99.39	99.40	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.49	99.50
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.87	26.69	26.60	26.50	26.41	26.32	26.22	26.13
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.50	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23	2.13
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.10
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.03
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
.....	6.63	4.61	9.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

Appendix Table 6 : Random Numbers

	1-4	5-8	9-12	13-16	17-20	21-24	25-28	29-32	33-36	37-40
1	2315	7548	5901	8372	5993	7624	9708	8695	2303	6744
2	0554	5550	4310	5374	3508	9061	1837	4410	9622	1343
3	1487	1603	5082	4043	6223	5005	1003	2211	5438	0834
4	3897	6749	5194	0517	5833	7880	5901	9432	4287	1695
5	9731	2617	1899	7553	0870	9425	1258	4154	8821	0513
6	1174	2693	8144	3393	0872	3279	7331	1822	6470	6850
7	4336	1288	5911	0164	5623	9300	9004	9943	6407	4036
8	9380	6204	7838	2680	4491	5565	1189	3258	4755	2571
9	4954	0131	8108	4298	4187	6953	8296	6177	7380	9527
10	3676	8726	3337	9482	9569	4195	9686	7045	2748	3880
11	0709	2523	9224	6271	2607	0655	8453	4467	3384	5320
12	4331	0010	8144	8638	0307	5255	5161	4889	7429	4647
13	6157	0053	6006	1736	3775	6314	8951	2335	0174	6993
14	3135	2837	9910	7791	8941	3157	9764	4862	5848	6919
15	5704	8865	2627	7959	3682	9052	9565	4635	0653	2254
16	0924	3442	0068	7210	7137	3072	9757	5609	2982	7650
17	9795	5350	1840	8948	8329	5223	0825	2122	5326	1587
18	9373	2595	7043	7819	8885	5667	1668	3695	9964	4569
19	7262	1112	2500	9226	8264	3566	6594	3471	6875	1867
20	6102	0744	1845	3712	0794	9511	7378	6699	5361	9378
21	9783	9854	7433	0559	1718	4547	3541	4422	0342	3000
22	8916	0971	9222	2329	0637	3505	5454	8988	4381	6361
23	2596	6882	2062	8717	9265	0292	3528	6248	9195	4883
24	8144	2317	1905	0495	4806	7569	0075	6765	0171	6545
25	1132	2549	3142	3623	4386	0862	4976	6762	2452	3245

মানুষের জ্ঞান ও ভাবকে বইয়ের মধ্যে সঞ্চিত করিবার যে একটা প্রচুর সুবিধা আছে, সে কথা কেহই অস্বীকার করিতে পারে না। কিন্তু সেই সুবিধার দ্বারা মনের স্বাভাবিক শক্তিকে একেবারে আচ্ছন্ন করিয়া ফেলিলে বুদ্ধিকে বাবু করিয়া তোলা হয়।

—রবীন্দ্রনাথ ঠাকুর

ভারতের একটা mission আছে, একটা গৌরবময় ভবিষ্যৎ আছে, সেই ভবিষ্যৎ ভারতের উত্তরাধিকারী আমরাই। নূতন ভারতের মুক্তির ইতিহাস আমরাই রচনা করছি এবং করব। এই বিশ্বাস আছে বলেই আমরা সব দুঃখ কষ্ট সহ্য করতে পারি, অন্ধকারময় বর্তমানকে অগ্রাহ্য করতে পারি, বাস্তবের নিষ্ঠুর সত্যগুলি আদর্শের কঠিন আঘাতে ধূলিসাৎ করতে পারি।

—সুভাষচন্দ্র বসু

Any system of education which ignores Indian conditions, requirements, history and sociology is too unscientific to commend itself to any rational support.

—Subhas Chandra Bose

Price : Rs. 150.00

(Not for sale to the Student of NSOU)

Published by Netaji Subhas Open University, DD-26, Sector-I, Salt Lake, Kolkata - 700064 & Printed at Gita Printers, 51A, Jhamapukur Lane, Kolkata-700 009.