# UNIT 1: INTRODUCTION

## Contents


1.0 Aims and Objectives

1.1 Introduction

1.2 Statistics Defined

1.3 Importance of Statistics

1.4 Types of Statistics

1.4.1 Descriptive Statistics

1.4.2 Inferential Statistics

1.5 Model Examination Questions


## 1.0 AIMS AND OBJECTIVES

This unit will introduce you to statistics and its uses and importance. After completing the unit you will be able to:

- define statistics
- identify the types of statistics
- know the benefits of managerial statistics.

## 1.1 INTRODUCTION

Governments, businesses, researchers and scientists in the Natural or Social science need information for their activities. Most of these information requirements are quantitative and need a scientific approach or technique to gather and use.

## 1.2 STATISTICS DEFINED

The world statistics is an Italian word composed of two words, stato, which means the state and statista-refers to a person involved with the affairs of the state. Therefore statistics was meant the collection of facts useful to the state.

Nowadays statistics in not restricted to information about the state. It extends to almost every realm of human endeavor.

Statistics is defined as a science or process of collecting, organizing, presenting, analyzing and interpreting data to assist in making effective decision.

## 1.3 IMPORTANCE OF STATISTICS

Statistics is useful for:

- Government officials for making policy decisions in unemployment, inflation, health, education, infrastructure etc…
- Financial planners for trend analysis, stock market, future investment etc..
- Businesses, for product development, customer satisfaction, Risk
- Production supervisors for quality control, improve product quality etc.
- Politicians for legislation campaign strategy
- Physicians and Hospitals on effectiveness of drugs and disease surveillance etc.

Managerial statistical analysis of data used to help in improving business processes to.

1- Demonstrate the need for improvements
2- Identity ways to make improvements
3- Asses weather or not improvement activities have been successful and
4- Estimate the benefits of improvement strategies

Statistical methods are used for learning about population, which is a set of existing units (people, objects or events)

Often the population that we want to study is very large, time consuming or costly to conduct a census. In such a situation we select and analyze a

subset (or portion) of the population units. This subset of the units in a population is called sample.

## 1.4 TYPE OF STATISTICS

**There are tow types of Statistics**

### 1.4.1 Descriptive Statistics

It is the science of describing the important aspects of a set of measurements eg. If we are studying a set of starting salaries we might wish to describe.

- How large or small they tend to be
- What a typical Salary should be
- How much the salaries differ from each other

When the population of interest is small and we can conduct a census of the population we will be able to directly describe the important aspects of the population measurement. The subject area of descriptive statistics includes procedures used to summarize masses of data and present them in an understandable manner. However it has nothing to do with the future.

### 1.4.2 Inferential Statistics

A Conclusion drawn about a population based on information in a sample drawn from the population is called statistical inference. Statistics is usually concerned with inference. The population we want to study is usually large or infinite. So we need to select a sample since it is impossible to study the population.

## 1.5 MODEL EXAMINATION QUESTIONS

Answer the following questions. Do not look into the text while writing the answers. However at the end refer to the text and see how you answered the questions.

a) Why governments, businesses, researchers need information?
b) Define statistics.
c) What are the types of statistics?
d) What are the particular benefits or importance of managerial statistics in improving business processes?

## UNIT 2:PROBABILITY AND PROBABILITY DISTRIBUTION

## Contents


2.0 Aims and Objectives

2.1 Introduction

2.2 Probability Defined

2.3 Approaches in Probability

- 2.3.1 Objective Probability
  - 2.3.1.1 Classic probability
  - 2.3.1.2 Long-term Relative Frequency Probability
- 2.3.2 Subjective Probability

2.4 Sample Space and Sample Space Outcome

2.5 Probability Rule

- 2.5.1 Addition Rule for Independent Events
- 2.5.2 Addition Rule for Mutually Exclusive Events

2.6 Complement of an Event

2.7 Conditional Probability and Statistical Independence

- 2.7.1 Conditional Probability
- 2.7.2 Statistical Independence
- 2.7.3 Independent and Mutually Exclusive Events
  - 2.7.3.1 Multiplication Rule for independent Events
  - 2.7.3.2 Union Rule for Independent Events

2.8 The Total Probability and Baye's Theorem

- 2.8.1 Total Probability
- 2.8.2 Baye's Theorem

2.9 Answers to Check Your progress

2.10 Model Examination Questions

## 2.0 AIMS AND OBJECTIVES

Probability theory forms the basis for inferential statistics as well as other fields that require quantitative assessment of chance occurrences; such as quality control, management decision analysis; and in areas of the natural sciences, engineering, economics etc.

After completing this unit, you will be able to

- define probability
- define important terms in probability
- identify the approaches in probability
- list sample space of an experiment
- identify the types of events
- calculate probabilities using deferent rules.

## 2.1 INTRODUCTION

Since life is full of uncertainties, people have always been interest in evaluating probabilities. The theory of probability is an in indispensable tool in the analysis of situations involving uncertainty.

## 2.2 PROBABILITY DEFINED

Probability can be defined as

- A mathematical means of studying uncertainty and variability.
- A number that conveys the strength of our belief in the occurrence of an uncertain event

From the above definitions you can differentiate probability to chances or possibilities. As the latter cannot be quantified.

Probability is a number between zero and one inclusive. The probability of zero represents something that cannot happen and the probability of one represents something that is certain to happen. The closer a

probability is to zero, the more improbable it is that something will happen the closer the probability is to one the more sure we are it will happen. When probability is 0.5 uncertainty will reach its maximum.

### *Important Terms*

**1. Experiment**

A process that leads to the occurrence of one and only one of several possible observations or

A process of observation that has an uncertain outcome. eg Tossing a coin; answering a question where the answer can be correct or incorrect; drawing a card from a deck of playing card.

**2. Event**

A collection of one or more outcomes of an experiment or

An experimental outcome that may or may not occur. If the experiment is tossing a coin the events are Head, or Tail.

**3. Outcome**

A particular result of an experiment. In case of tossing a coin, If head face up we will consider head as the out come of the experiment.

## 2.3 APPROACHES IN PROBABILITY

### 2.3.1Objective Probability

#### *2.3.3.1 Classic Probability*

It is probability based on the symmetry of games of chance or similar situations. This probability is based on the idea that certain occurrences are equally likely. eg. The numbers 1,2,3,4,5,and 6 on fair die are equally likely to occur i.e they do have equal chance of occurrence.

#### *2.3.1.2 Long-term Relative Frequency Probability*

The probability of an event happening in the long-term is determined by observing what fraction of the time similar events happened in the past. We often think of a probability in terms of the percentage of the time the event would occur in many repetition of the experiment. Suppose that A

is an event that might occur when a particular experiment is performed then the probability that the event A will occur, P (A), can be interpreted to be the number that would be approached by the relative frequency of the event A If we perform the experiment an indefinitely large number of times.

eg. When we say that the probability of obtaining a head when we toss a coin is 0.5 we are saying that, when we repeatedly toss the coin an indefinitely large number of times, we will obtain a head 50% of the repetition.

In terms of formula

$$\text{Probability of an event happening} = \frac{\text{Number of times occurred in past}}{\text{Total number of observation}}$$

If a truck operator experienced 5 accidents out of 50 truck last year, then the probability that a truck will have an accident next year can be 5/50 = 0.10

### 2.3.2 Subjective Probability

When there is no past experience or little on which to base a probability, personal judgment, experience, intuition or expertise or any other subjective evaluation criteria will be applied to estimating or assigning probability. This probability is subjective probability.

It is also called personal probability. Unlike objective probability one person's subjective probability may very well different from another person's subjective probability of the same event.

eg. A physician assessing the probability of a patient's recovery and an expert in the national bank assessing probability of currency devaluation are both making a personal judgment based on what they know and feel about the situation and other group of physicians or experts will arrive

with different probability, though both can employee identical techniques or approaches and information.

Both classic and long-term relative frequency probabilities are objective in the sense that no personal judgment is involved.

Whatever the kind of probability involved /subjective or objective/ the same set of mathematical rules holds for manipulating and analyzing probability.

## 2.4 SAMPLE SPACE AND SAMPLE SPACE OUTCOME

In order to calculate and interpreter probabilities it is important to understand and use the idea of sample space.

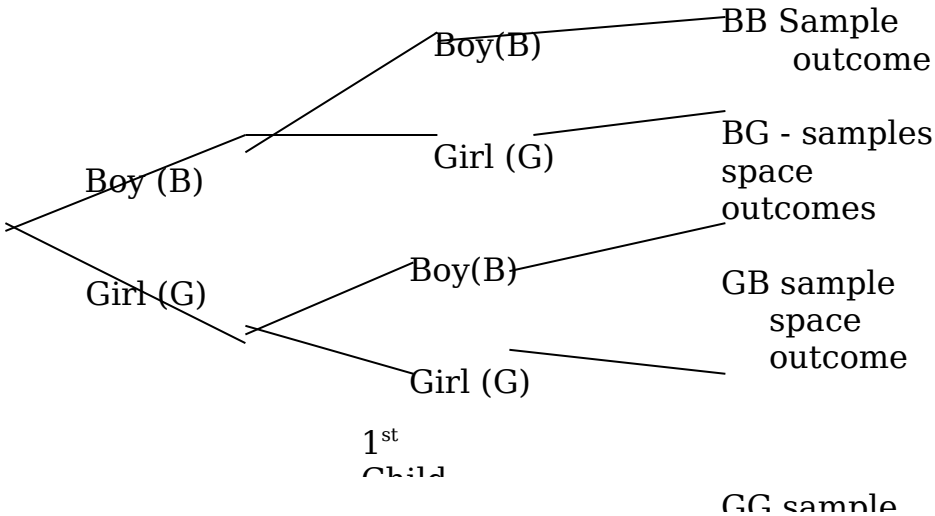
*The sample space of an experiment* is the set of all of the distinct possible outcomes of the experiment. Each distinct out come is called sample space out come or sample point or elementary event.

**Example 1**

A newly married couple plans to have two children. Naturally, they are curious about whether their children will be boys or girls. Therefore, we consider the experiment of having two children.

In order to find the sample spaces of this experiment, of having two children, we let ‘B’ denote that child is a boy and ‘G’ denotes that child is a girl.

This experiment is a two-step process i.e having the first child, which could be a boy or a girl and having the second child, which could also be either a boy or a girl.
This can be constructed by a tree diagram. Each branch of the tree leads us to a distinct sample space outcome.




We see that there are four sample space outcomes. Therefore the sample space (i.e the set of all of the distinct samples space outcomes is BB BG GB GG.

In order to consider the probabilities of these outcomes, suppose that boys and girls are equality likely each time a child is born. This says that each of the sample space out comes is equally likely. i.e.
P(BB) = p(BG)=p(GB)=p(GG)= $^1/_4$ This says that there is a 25%, chance that each of these outcomes will occur. Since we are certain that there is no other option or combination remaining, the probability that the couple will have any one of the sample space outcomes is one. i.e. P(BB) + P(BG) + P(EB) + P(EG) = 1

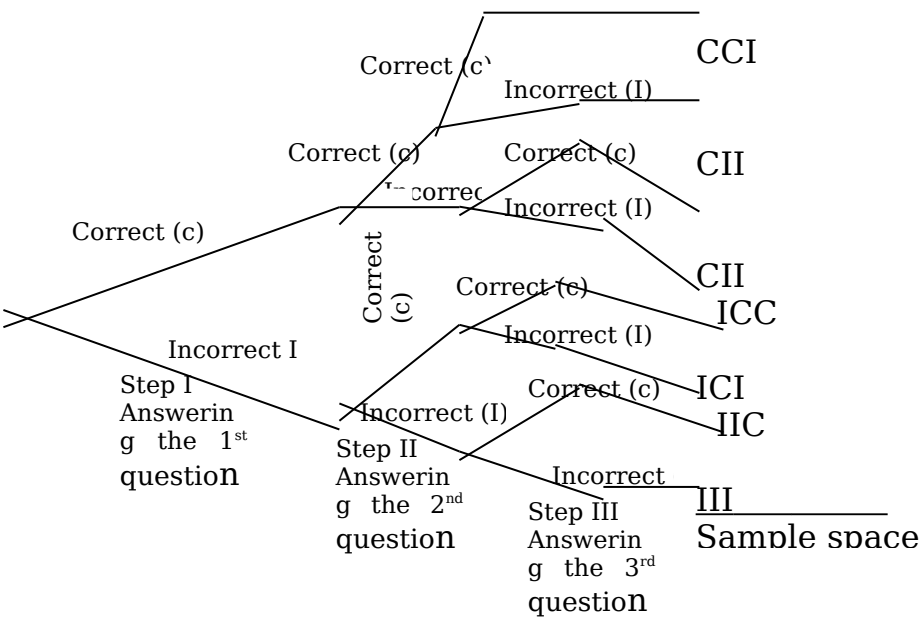
Notice that these probabilities sum one i.e the sum of the probabilities of all sample space outcomes is one.

Therefore the sample space (that is, the set of all of the distinct sample space out comes) is BB, BG, GB, GG

**Example 2**

A student takes a quiz that consist of three true or false questions. If we consider our experiment to be answering the three questions, each question can be answered correctly or incorrectly.

Let c denote answering a question correctly and I denote answering a question incorrectly. Then we can depict a tree diagram of the sample space out come for the experiment.




This diagram portrays the experiment as a three-step process

Step I – answering the 1st question (Correctly or incorrectly) (C or I)

Step II – answering the 2nd question (Correctly or incorrectly).

Step III – answering the 3rd question (Correctly or incorrectly).

The tree diagram has eight different branches and the eight distinct sample space outcomes are listed at the end of the branches. We see the sample space is

CCC CCI CIC CII

ICC ICI IIC III

Now suppose that the student was totally unprepared for the test, and has to blindly guess the answer to each question that is the student has a

50-50 chance or 0.5 probability of correctly answering each question. This means that each of the eight sample space outcomes is equally likely to occur.

i.e

P(ccc) = P(ccI) ------P(III) =1/8

Here also the sum of the probabilities of the sample space out comes is one.

In General the sum of the probabilities of all the sample space is equal to 1.

**Finding Probabilities by using Sample Space**

If all of the sample space out comes are equally likely, then the probability that an event will occur is equal to the ratio:

**The number of sample space outcomes that correspond to the event**

The total number of sample space outcomes.

Consider the couple planning to have two children to find the probability of two boys first we have to find the sample space outcome corresponding to the event of having the first child a boy and the second child also a boy.

There is only one sample space outcome corresponding to this event i.e. BB so the probability will be: $\frac{1}{4}$= 0.25 the probability that the couple will have a boy and a girls is similarly calculated by first identifying the sample space outcomes corresponding to the event of having a boy and a girls. The sample space outcomes are BG and GB. So the probability will be $\frac{2}{4}$= 0.5

## Check Your Progress – 1

1. Suppose that a couple will have three children. Letting B denote a boy and G denote a girl.

a) Draw a tree diagram depicting the sample space out come for this
experiment.

b) list the sample space outcomes that correspond to each of the following events.
    1) All three children will have the same gender
    2) Exactly two of the three children will be girls.
    3) Exactly one of the three children will be a girl.
    4) None of the tree children will be a girl.

2. Four people will enter an automobile show Room and each will either purchase a car (P) or will not purchase a car (N)
    a) Draw a tree diagram depicting the sample space of all possible purchase decision that could potentially be made by the four people.
    b) List the sample space out comes that correspond to each of the following events.
        1) Exactly three people will purchase a car
        2) Two or fewer will purchase a car
        3) One or more people will purchase a car
        4) All four people will make the same purchase decision

Often time it may be practically impossible to list all possible sample space outcomes of an experiment. Under such circumstances we can find the probability of an event by identifying the number of sample space outcomes /without listing/ corresponding to the event.

**Example** - Suppose that 650.000 of 1,000,000 households in Addis subscribe to a newspaper called Addis Zemen, and consider randomly selecting one of the Households in this city. That is consider selecting one

household & giving each and every household in the city the same chance of being selected. Let A be the event that the randomly selected household subscribes to the Addis Zemen. Then since the sample space of this experiment consists of 1,000,000 equally likely sample space outcomes (households). It follows that

P(A) = the number of Households that subscribe to the Addis Zemen / The total number of households in the city

= 650,000 / 1000,000 = 0.65

Now also suppose that 500,000 households in the city subscribe to the Ethiopian Herald (H) and further suppose that 250,000 households subscribe to both the newspapers.

We consider randomly selecting one household in the city, and we define the following events

A = The random of selected house hold subscribes to the Addis Zemen.

Ā = The randomly selected, hose hold does not subscribe to the Addis Zemen.

H = The randomly selected household subscribes to the Ethiopian Herald.

$\overline{H}$ = The randomly selected household does not subscribe the Herald.

Using the notation AnH to denote both A& H we also define.

AnH = The randomly selected household subscribes both to Addis Zemen & Herald.

Since 650,000 of the 1,000,0000 households subscribe to the Addis Zemen (that is correspond to the event Occurring). Then 350,000 households do not subscribe to Zemen (Ā) i.e. 1,000,000 – 650,000.

Similarly since 500,000 households subscribe to Herald (H) 500,000 households do not subscribe to herald ($\overline{H}$).

Next consider the events

An$\overline{H}$ = the randomly selected household subscribes to Zemen and does not subscribe to Herald;

ĀnH = the randomly selected household does not subscribe to Zemen and does subscribe to Herald.

**A summary of the number of house holds corresponding to the events A, Ā, H, $\overline{H}$ and AnH**

| Events | Subscribe to Herald | Does not subscribe to Herald | Total |
|---|---|---|---|
| Subscribe & Addis Zemen | 250,000 | | 650,000 |
| Does not subscribe to Addis Zemen | | | 350,000 |
| Total | 500,000 | 500,000 | 1,000,000 |

Define the event Ā n$\overline{H}$,

Ā n$\overline{H}$ = the randomly selected household does not subscribe to both newspaper.

Since 650,000 households subscribe to the Addis Zemen (A) and 250,000 households subscribe to both Zemen and Herald (AnH) it follows that 650,000 – 250,000 = 40,000 house holds subscribe to Addis Zemen but do not subscribe to Herald, (An$\overline{H}$). This subtraction is illustrated in the table below.

By similar logic

a. 500,000 – 250,000 = 25,000 households do not subscribe to Addis Zemen but do subscrige to Herald (Ā nH)
b. 350,000 – 250,000 = 100,000 households do not subscribe the Addis Zemen and also do not subscribe the Herald (Ā n$\overline{H}$)

c. Subtracting to find the number of households corresponding to the events.

d. AnH, An $\overline{H}$,

| Event | H | $\overline{H}$ | |
|---|---|---|---|
| A | 250,000 | 650,000-250,000 | 650,000 |
| Ā | | | 350,000 |
| Total | 500,000 | 500,000 | 1,000,000 |

e. (Ā n H) = 5000,000-250,000

= 250,000

f(Ā n $\overline{H}$) = 350,000 – 250,000

= 100,000

A contingency table summarizing subscription data for Addis Zemen and Herald

| Event | Subscribe to Herald (H) | Does not Subscribe to Herald ($\overline{H}$) | Total |
|---|---|---|---|
| Subscribe to Addis Zemen (A) | 250,000 | 400,000 | 650,000 |
| Does not subscribe to Addis Zemen (Ā) | 250,000 | 100,0000 | 350,000 |
| Total | 500,000 | 500,000 | 1,000,000 |

Now since we will randomly select one household (making all the households equally likely to be chosen), the probability of any of the previously defined events is the ration of the number of households corresponding to the event's occurrence to the total number of households in the city.

Therefore

$$P(A) = \frac{650,000}{1,000,000} = 0.65$$

$$P(H) = \frac{500,000}{1,000,000} = 0.5$$

$$P(AnH) = \frac{250,000}{1,000,000} = 0.25$$

Next letting AUH denote either A or H, we consider finding the probability of the event

AUH = the randomly selected household subscribes to either the Addis Zemen or Herald. (i.e subscribe to at least one of the two newspapers).

We see that the households subscribing to either Addis Zemen or Herald:

a) The 400,000 households that subscribe to only Addis Zemen, An $\overline{H}$
b) The 250,000 house holds that subscribe to only the Herald, ĀnH and
c) The 250,000 households that subscribes to both Addis Zemen and Herald, AnH. Therefore since a total of 900,000 households subscribe to either the Addis Zemen or Herald it follows: -

$$P(AUH) = \frac{900,000}{1,000,000} = 0.9$$

i.e 90% of the house holds in the city subscribe to either Addis Zemen or Herald.

Notice that P(AUH) = 0.9 does not equal

P(A) +P(H) = 0.65 +0.5 = 1.15

Logically the reason for this is that both P(A) = 0.65 and P(H) = 0.5 count the 25% of the households that subscribe to both newspapers. Therefore;

the sum of P(A) and P(H) counts this 25% of the households once to often

It follows that if we subtract P(AnH) = 0.25 from the sum of P(A) and P(H) then we will obtain P(AuH) i.e

P(AuH) = P(A)+P(H) – P(AnH)

= 0.65+0.5-0.25 = 0.90

**The intersection and union of Two events.**

Given two events A&B

1) The Intersection of A&B is the event consisting of the sample space outcomes belonging to both A&B, denoted AnB. Further more P(AnB) denotes the probability that Both A&B will simultaneously Occur.
2) The union of A&B is the event consisting of sample space outcomes belonging to either A or B. The union is denoted AUB Further more P(AUB) denotes the probability that either A or B will occur.

## 2.5 PROBABILITY RULES

### 2.5.1 The Addition Rule

#### 2.5.1.1 Addition Rule for two Dependent Events

Let A and B be events then the probability that either A or B will occur is

P(AUB) = P(A)+P(B)-P(AnB)

#### 2.5.1.2 Addition Rule for Two Mutually Exclusive Events

Two events are said mutually exclusive if they have no sample space outcomes in common. In this case the event A&B cannot occur simultaneously and thus.

P(AnB) = 0

Let A&B Mutually exclusive events then, the probability that either A or B will occur is

P(AUB) = P(A) + P(B)

**Example** - consider randomly selecting a card from a standard deck of 52 playing cards and define the events.

J, a randomly drawn card is Jack; Q, a randomly drawn card is Queen; and K, a randomly drawn card is a king.

Since there are 4 Jacks, 4 Queens and 4 Kings in the deck.

P(Q) = $4/52$ P(K) = $4/52$ P(J) = $4/52$

Since there is no card that is both a J & Q the event J and Q are mutually exclusive and thus P(JnQ) = 0 it follows that the probability that the randomly selected card is either J or Q is

P(JUQ) = P(T) + PQ

= 4/52 + 4/52 = 2/13

### 2.5.1.3 The Addition Rule for N mutually exclusive events.

The event $A_1$, $A_2$, ------$A_n$ are mutually exclusive if no two of the events have any sample space out come in common. In this case no two of the events can occur simultaneously and

$P(A_1UA_2U\text{-----}UA_n) = P(A_1)+P(A_2)+\text{-----}+P(A_n)$

**Example** P(JuQUKU nine) =

P(J)+P(Q) +P(K) +P(nine)

= $4/52$ + $4/52$ + $4/52$ + $4/52$ = $\frac{16}{52}$

## 2.6 THE COMPLEMENT OF AN EVENT

Given an event A, the complement of A is the event consisting of all sample space outcomes that do not correspond to the occurrence of A.

The complement of A is denoted $\bar{A}$

Furthermore $P(\bar{A})$ denotes the probability that A will not occur.

In any probability situation, either an event A or its compliment A must occur.

Therefore we have

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

This implies

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

**Example** – If team A and B are playing for a final cup we can say that the events that team. A will win is complement to the event that B will win. i.e., if A wins B will lose. Under no circumstance that A will win and looses at the same time winning and losing are mutually exclusive.

## 2.7 CONDITIONAL PROBABILITY AND INDEPENDENCE

### 2.7.1 Conditional Probability

Probability is conditional upon information. We may define the probability of event A conditional upon the occurrence of event B.

If we think about two adjacent rooms, $R_1$ and $R_2$, the probability that R1 will be caught by fire is highly conditional on the probability of the other room.

**Example 1**. Suppose that we randomly select a household, and that the chosen house hold reports it subscribes to Herald. Given this new information we wish to find the probability that this household subscribes to Addis Zemen. The new probability is called a conditional probability.

The probability of the event A, given the condition that the event H has occurred, is written

P(A/H) = the probability of A given H. We often refer to such a probability as the conditional probability of A given H .

In order to find the conditional probability that a household subscribes to Addis Zemen given that it subscribes to Herald we know that we are considering one of 500,000 households. Since 250,000 of these 500,000 Herald subscribers also subscribe to Addis Zemen we have

$$P(A/H/ = \frac{250{,}000}{500{,}000} = 0.5$$

i.e 50% of the Herald subscribers also subscribe to Addis Zemen:

**Example 2.** Next suppose that we randomly select another household from the 1,000,000 house holds and suppose that this newly chosen household reports that it subscribes to Addis Zemen

Now find the probability that this house hold subscribes to Herald

P(H/A)

$= \frac{250,000}{650,000} = 0.3846$

This says that the probability that the randomly selected household subscribes to herald given that the household subscribes to Addis Zemen is 0.3846. ie., 38.46% of Addis Zemen subscribers also subscribe to Herald.

We have

$P(A) = \frac{650,000}{1,000,000} = 0.65$

$P(AnH) = \frac{250,000}{1,000,000} = 0.25$

$P(H/A) = \frac{250,000}{650,000} = 0.3846$

$P(H) = \frac{500,000}{1,000,000} = 0.5$

$P(A/H) = \frac{250,000}{500,000} = 0.5$

If we divide both the numerator and denominator of each conditional probability by 1,000,000

$P(A/H) = \frac{250,000}{500,000} = \frac{250,000/1,000,000}{500,000/1,000,000} = \frac{P(AnH)}{P(H)}$

$P(H/A) = \frac{250,000}{650,000} = \frac{250,000/1,000,000}{650,000/1,000,000} = \frac{P(AnH)}{P(A)}$

We express these conditional probabilities in terms of P(A), P(H) and P(AnH)

Given the sample space outcomes are equally likely.

$P(A/H) = \frac{P(AnH)}{P(H)}$ , then P(AnH) = P(H) P(A/H), by simple cross multiplication

$P(H/A) = \frac{P(AnH)}{P(A)}$ = then P(AnH) = P(A) P(H/A)

**The General Multiplication Rule**

(two ways to calculate P(AnH)

$$P(AnH) = P(A)\,P(H/A) = P(H)\,P(A/H)$$

Example 1. In a firm 20% of the employees have an accounting background, while 5% of the employees are executives and have an accounting backgrounds. If an employee has accounting background, what is the probability that the employee is an executive.

Let us define the events

E, an employee is an executive and

A, an employee has an accounting background

P(A) = 0.2

P(AnE) = 0.05

then

$$P(E/A) = \frac{P(AnE)}{P(A)} = \frac{0.05}{0.2} = 0.25$$

**Example 2**. A contractor is bidding for two projects with Co. A and Co. B. The contractor estimates that the probability of obtaining the project with Co. A is 0.45. He also fells that if he should get the project with Co. A then there is a 0.90 probability that Co. B will also give him the project. What are the contractor's chances of getting both projects:

**Solution:** We are given

P(A) = 0.45

P(B/A) = 0.90 and we are looking for P(AnB), which is the probability that both A and B will occur. From the equation we have
P(AnB) = P(B/A) P(A) = 0.9 x 0.45 = 0.405

## Check Your Progress –2

21% of the executive in a large firm are at the top salary level. It is further known that 40% of all the executives at the firm are women. Also 6.4% of all executives are women and are at the top salary level. Recently among executives at the firm arose a question as to whether there is any evidence of salary inequality. Check.

Clue. To solve this problem, pose a question in terms of probabilities. I.e., ask whether the probability that the executive will be at the top salary level given the executive is a woman. If this probability is less than 16% (the average) you can conclude that salary inequity does exist because of gender.

### 2.7.2 Statistical Independence

If the occurrence of events A and B have nothing to do with each other, then we know that A and B are independent events.

i.e the probability of occurrence of A well not influence the probability of occurrence of B.

This implies that

P(A/B)= p(A) and that

P(B/A) = p(B)

Further more the general multiplication rule tells us that, for any two events A and B we can say that

P(A n B) = p(A) p(B/A) there fore if p(B/A)= p(B) if follows that

P(AnB) = p(A) p(B)

This is called the multiplication rule for two independent events.

However, if the probability of an event is influenced by whether or not another event occurs, we say the two events are dependent.

eg. Define the events C and P as follows

C= your favorite college football team will win its first match next season.

P= Your favorite professional football team will win its first match next season.

Suppose that you believe that for next season p(c) = 0.6 and p(p) =0.6 then since the outcomes of a college football games and a professional football game would probably have noting to do with each other, it is reasonable to assume that C and P are independent events.

It follows that : Both your favorite teams will win their first match next season,

P(CnP)= p(c)  p(p)=0.6(0.6)=0.36

When two events are independent, neither are their complements.

### 2.7.3 Independent and Mutually Exclusive Events

When two events are mutually exclusive they are not independent. In fact they are very dependent events in the sense that if one happens the other cannot happen. The intersection of two mutually exclusive events is zero but the probability of the intersection of two independent events is not zero. It is equal to the product of the probabilities of the separate events.

### 2.7.3.1 The multiplication rule for N independent events

The events $A_1$, $A_2$ …. An are independent events if the occurrence of these events have nothing to do with each other. if events $A_1$, $A_2$, …, $A_n$ are independent events, then

$P(A_1 \text{ n} A_2 \text{ n}. . . \text{n} A_n)= P(A_1) P(A_2). . . P(A_n)$

**Example 1**.  An electronic devise has four independent components $C_1$, $C_2$, $C_3$, C4, with a reliability of 0.85 each. The device works only if all four components are functional.

What is the probability that the device will work when needed?

P(the device will work) = P(all components will work ) = P($c_1$, $nc_2$,$nc_3$,$nc_4$)

= p($C_1$) p($C_2$) p($C_3$) p($C_4$)

= 0.85 x 0.85 x 0.85 x 0.85

=0.85 x 0.85 x 0.85 x 0.85 = 0.522

**Example 2.** The rate of defects in corks of wine is 0.75. Assuming independence, if four bottles are opened ($B_1$, $B_2$, $B_3$, $B_4$), what is probability that four corks are defective.

P(all 4 are defective)= P($B_1$ n $B_2$ n $B_3$ n $B_4$) = P($B_1$) P($B_2$) P($B_3$) P($B_4$)

= 0.75 x 0.75 x 0.75 x 0.75=0.316

### 2.7.3.2 Union rule

The union of several independent events is the event that at least one of the events happens.

The probability of the union of several independent events $A_l$, $A_2$, … $A_n$ is

P(A, u$A_2$ u. . . u$A_n$) = 1- P($\bar{A}_1$) P($\bar{A}_2$). . . p($\bar{A}_n$)

**Example 1:** A device similar to the above one has three components, but the device works as long as at least one of the components is functional. The reliability of the components are 0.96, 0.91 and 0.80what is the probability that the device will work when needed?

P(The device will work) = p(at least one will work) = 1 – p(all will fail)

=1–p($\overline{c_1}$)P($\overline{c_2}$)P($\overline{c_3}$) = 1-(0.04) (0.09)0.02

**Example 3:** In the developing world a woman's adds of dying from problems related to pregnancy is 1 in 51. If three women are pregnant what is the probability that at least one will die

p(at least one will die)= 1- p(all will survive)

$1-(50/51)^3 = 0.0577$

## 2.8 THE TOTAL PROBABILITY AND BYE'S THEOREM

### *2.8.1 Total Probability*

What ever may be the relationship between two events we can always say that the probability of A is equal to the probability of the intersection of A and B plus the probability of the intersection of A and the complement of B (event $^{-}$B)

$P(A) = P(AnB) + P(A\ n^{-}B)$

Total probability

Consider the households subscribing to the two newspapers.

$P(A) = 0.65$

This probability includes the households subscribing to both the newspapers P(AnB) or the households subscribing to Addis Zemen and not for Herald. I.e.,

$P(A) = P(AnN) + P(An^{-}H)$

$= 0.25 + 0.40$

$= \underline{\underline{0.65}}$

The law of total probability may be extended to more complex situations, where the sample space X is portioned into more then two events. Say we partition the sample space in to a collection of n sets $B_1$, $B_2$…$B_n$ The law of total probability in this situation is

$$P(A) = \sum_{i=1}^{n} p(AnB_i)$$

**Example 1:** Suppose A is the event that a picture card is drawn out of a standard deck of 52 cards Letting H.C.D and S denote the events that the card drawn is a Heart, Club, Diamond or Spade respectively.

In a standard deck there are 12 picture cards. The probability will then be $^{12}/_{52.}$ Following the law of total probability. This probability can be obtained as the sample of the intersections of the four events with A. In the deck there are three pictured cards and Heart (Jack heart, queen

hearts and king heart), three pictured and club; there pictured and diamond and three pictured and at the same time spade.

We find the probability of a picture card, P(A)

P(A) = P(AnH) + P(AnC) + P(AnD) +P(AnS)

$= {}^{3}/_{52} + {}^{3}/_{52} + {}^{3}/_{52} = {}^{12}/_{52}$

The law of total probability can be extended using the definition of conditional probability.

P(AnB) = P(A/B) p(B) similarly

P(An¯B) = P(A/¯B) p(¯B)

Substituting this formula to the addition

i.e P(A)= p(AnB) + p(An¯B)

P(A) = P(A/B) p(B) + p(A/¯B)(¯B)

For more than two sets

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

Where there are n sets in the partition

**Example 1:** An analyst believes that the market has a 0.75 probability of going up in the next year if the economy should do well, and a 0.30 probability of going up if the economy should not do well during the year. The analyst further believes there is a 0.80 probability that the economy will do well in the coming year.

What is the probability that the market will go up next year?

Define the events

U= The Market will go up

W= The economy will do well

Find p(U)

P(U/W) = 0.75

If P(W) = 0.8 then

P(W) = 0.80

P($\overline{W}$= 1 – 0.8 = 0.2

P(U/$\overline{W}$) = 0.3

P(u) = P(u/W)p(w) + p(u/¯w ) p(¯w

=0.75(080) + 03(0.2)

=0.66

This means the market can go up in two ways i.e if the economy will do well and the market will go up and if the economy will not do well and the market will go up.

## 2.8.2 Baye's Theorem

Baye's Theorem is a very important theorem to revise probabilities using some additional information. First let us define to important terms.

### 2.8.2.1 Prior Probability /Initial Probability)

It is a given probability before any empirical data is observed

### 2.8.2.2 Posterior Probability

Is revised probability based on new information. Prior probabilities can be reviized as we have additional or new information about the events.

Derivation

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

By another definition

$$\text{i.e., } P(AnB) = \frac{P(A/B)\ P(B)}{P(A)}$$

From the law of total probability

$$P(A) = P(A/B)\ P(B) + p(A/^{-}B)\ P(^{-}B)$$

Substituting this expression for P(A) in the denominator

$$P(B/A) = \frac{P(A/B)\ p(B)}{P(A/B)\ p(B)+p(A/^{-}B)p(^{-}B)}$$ Baye's Theorem

The probabilities p(B) and p($^{-}$B) are called prior probabilities of the events B and$^{-}$B. The probability P(B/A) is called the posterior probability of B.

The theorem allows us to reverse the conditional it of events.

We can obtain the probability of B given A from the probability of A given B.

Baye's theorem may be viewed as a means of transforming one prior probability of an event B into a posterior probability of the event B posterior to the known occurrence of event A.

**Example1**. Let A be the event that a randomly selected American has the deadly disease AIDS. And let $\bar{A}$ be the event that the randomly selected American does not have AIDS. Since it is estimated that 0.6 percent of the American population have AIDS.

$P(A) = 0.006$ and $P(\bar{A})=0.994$

There is a test that attempts to detect whether a person has AIDS. According to historical data 99.9% of people with AIDS react positively (RP) to the test.

i.e $P(RA/A)=0.999$

Further more 1% of people with out AIDS react positively.

i.e., $P(RP/\bar{A}) = 0.01$

If we give a randomly selected American the test and the person reacts positively, what is the probability that the person actually has Aids?

The idea of Bay'es theorem is that we can find P(A/RP) by thinking as follows. A person will react positively (RP) if the person react positively and actually has AIDS (AnRP) or if the person react positively and does not actually have AIDS.

($\bar{A}$ nRP)

Therefore,

$P(RP) = P(AnRP) + P(\bar{A}\ nRP)$

This implies that

$$P(A/RP)= \frac{P(AnRP)}{P(RP)}$$

$$= \underline{P(AnRP)}$$

$$P(AnRP) + p(\bar{}AnRP)$$

$$= \frac{P(A)\ P(RP/A)}{P(A)\ p(RP/A) + P(\bar{}A)\ P(RP/\bar{}A)}$$

$$= \frac{(0.006)\ (0.999)}{(0.006)(0.999)+(0.994)(0.01)}$$

$$= \underline{\underline{0.38}}$$

This probability says that, if all Americans were given an AIDS test only 38%of the people who would react positively to the test would actually have AIDS.

Bay'es theorem may be extended to a partition of more than two sets. This is done using the law of total probability involving a partition of sets $B_1, B_2, \ldots . B_n$.

The theorem gives the probability of one of the sets in the partition B, Given the occurrence of event A.

Extended Bayes' theorem.

$$P(B_1/A) = \frac{P(A/B_1)P(B_1)}{\sum_{i=1}^{n} P(A/B_i)P(B_i)}$$

**Example 1.** An Economist believes that during periods of high economic growth the U.S dollar appreciates with probability 0.70; in periods of moderate economic growth the dollar appreciates with probability 0.40; and during periods of low economic growth the dollar appreciates with probability 0.20. During any period of time the probability of high economic growth is 0.30, the probability of moderate growth is 0.50 and the probability of low economic growth is 0.2. Suppose the dollar has been appreciating during the present period. What is the probability that the economy is experiencing a period of high growth. Define the three events,

High economic growth (H)

Moderate economic growth(M)

Low economic growth (L)

The prior probabilities of the three states of the economy are P(H) =0.3 P(M)= 0.5 P(L)=0.2 Let A denote the event that the dollar appreciate. We have the following conditional probabilities.

P(A/H)= 0.70 P(A/M) = 0.40 P(A/L)= 0.20

Find P(H/A)

$$= \frac{P(H/A) = P(A/H)\,P(H)}{P(A/H)\,P(H) \quad P(A/H)\,P(H) + P(A/M)+P(M)+P(M)+P(A/L)P(L) \quad P(AnH) + P(AnM) + P(AnL)}$$

$$= \frac{0.70(0.30)}{0.70(0.30) + 0.4(0.5) + 0.2(0.2)}$$

$$= \underline{0.467}$$

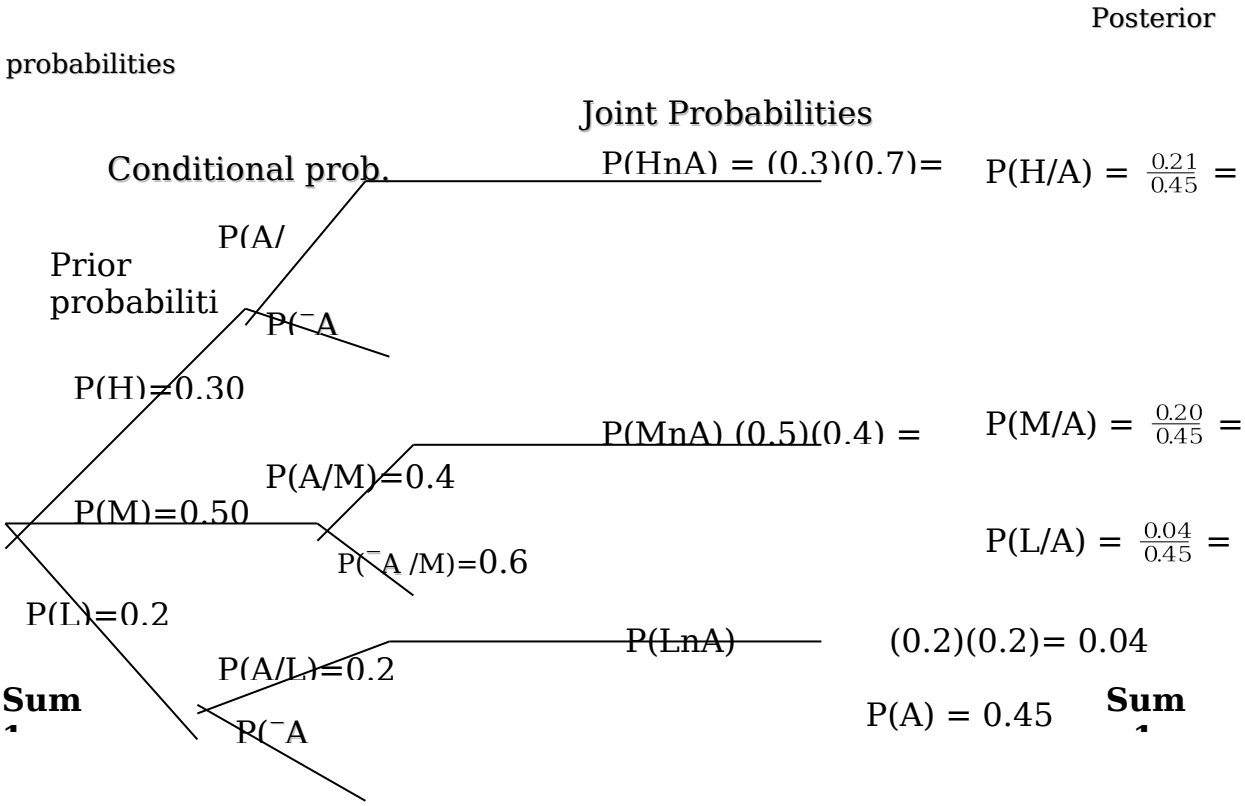
We can obtain this answer along with posterior probabilities of the other two states of the economy M and L. i.e P(M/A) and P(L/A)

| **Event** | **Prior probability** | **Conditional probability** | **Joint probability** | **Posterior probability** |
|---|---|---|---|---|
| H | P(H) 0.30 | P(A/H)=70 | P((AnH)=0.21 | P(H/A)=0.21/0.45=0.467 |
| | | | + | |
| M | P(M)=0.50 | P(A/M)=0.40 | P(AnM)=0.20 | P(M/A)=0.20/0.45= 0.444 |
| | | | + | |
| L | P(L)=0.20 | P(A/L) =0.2 | P(AnL)=0.04 | P(L/A)=0.04/0.45=0.089 |
| | | | = | |

Sum 1 P(A)=0.45 Sum =1

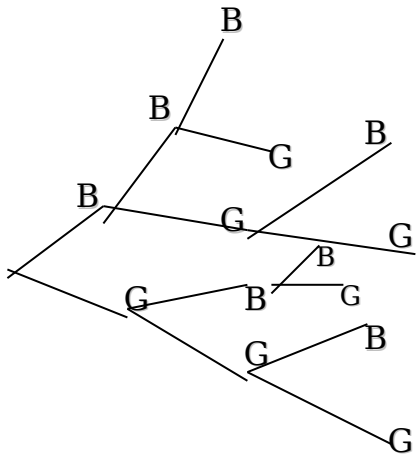
Note that both the prior probabilities and the posterior probabilities of the three states add to one.

Tree Diagram for the above example




## 2.9 ANSWERS TO CHECK YOUR PROGRESS

1)



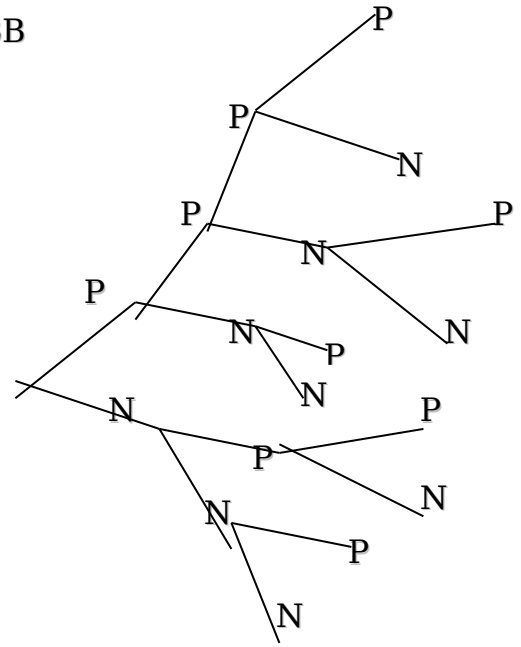
b) 1) B BB, GGG

2) GBG, BGG, GGB

3) BBG BGB GBB

4) BBB

2) a)




b) 1) PNPP, PPNP, PPPN, NPPP

2) NNNN, NNNP, NNPN, NPNN, PNNN, NNPP, NPNP, PNNP, PPNN, NPPN, PPNN

3) NNNP, NNPN, NPNN, PNNN, NNPP, NPNP, PNNP, PPNN, NPPN, PPNN, PNPP, PPNP, PPPN, NPPP, PPPP.

4) PPPP, NNNN

## 2.10 MODEL EXAMINATION QUESTION

**Part I. Define the following terms of words**

1. Probability
2. an experiment
3. an even
4. an outcome

5. objective probability
6. is subjective probability
7. sample space outcome
8. sample space
9. mutually exclusive events
10. Independent events
11. Dependent events
12. Complement of an event
13. Prior probabilities
14. Posterior probabilities.

**Part II. Workout the following questions**

Clearly show the steps

1. A newly established company is planning to recruit trainees for four jobs in the marketing department. The marketing manager contacted an employment agency. The agency has selected four candidates and send them to the Company.

   The company will hire those who fulfill the requirement of the job. Assuming that a candidate's chance to pass the final evaluation is 0.5.

   a. List all the sample space outcomes of the experiment
   b. Identity the sample space outcomes corresponding to the following events
      i. All of them will qualify
      ii. Only two of them will qualify
      iii. None of them will qualify
      iv. There of them will quality
   c. Assuming the probability that a candidate will be qualified for job is 0.5, find the probability for the events listed in part D.
2. The personnel manager of a company construct the following summary table about the efficiency of Company employees.

| Event | Efficiency | | | |
|---|---|---|---|---|
| | High, H | Average, A | Low, L | Total |
| Men | 120 | 100 | 80 | 300 |
| Women | 45 | 35 | 20 | 100 |
| Total | 165 | 135 | 100 | 400 |

Find the probability that a randomly selected employee

a) has high efficiency
b) has average efficiency
c) has low efficiency
d) has high efficiency given that this employee is
    i. a man
    ii. a woman
e) is a woman and has high efficiency
f) is a man and has low efficiency
g) has high or low efficiency

3. A firm is planning to introduce a new product. The probability that the product will be successful if a competitor does not come up with a similar product is 0.67. The probability that the new product will be successful in the presence of a competitor new product is 0.42. The probability that the competing firm will come out with a new product during the period is question is 0.35.
   What is the probability that the product will be a success?
4. 25% of college class graduated with honors, while 20% of the class were honors graduates and obtained good jobs. What is the probability that a person got a good job if he graduated with honors?
5. A contractor is bidding for four-construction project. He assesses his chances of winning the projects at 0.6, 0.75, 0.9 and 0.5. Assuming independence.
    a) What is the probability that the contractor will win all projects?

b) What is the probability that the contractor will win at least one project?

c) What is the probability that he will win none of the projects?

6. A package of documents needs to be sent to a given destination and it is important that it arrive with in one day. To maximize the chance of on time delivery, three copies of the document are sent via three different delivery services. Service A is known to have a 90% on time delivery record, service B has an 88% on time delivery record, and service C has 91% on time delivery record. What is the probability that at least one copy of the documents will arrive at its destination on time?
7. Three secretaries, $S_1$, $S_2$ and $S_3$ do office work for a company, mainly filling papers, of all the papers that come into the office, $S_1$ files 50% $S_2$ files 30% and $S_3$ files the rest. Each secretary occasionally misfiles a paper $S_1$ misfiles 5% of the papers she files, $S_2$ misfiles 7% of the papers she files and $S_3$ misfiles 10% of the papers she files. The manager has been looking for a particular paper and has found that it has been misfiled. He decides to give warning to the one who most likely filed it. Who most likely filed it? Draw a tree diagram.
8. A manufacturing Co. purchases a component form three different suppliers. When components arrive at the warehouse of the co. they are placed in a bin without inspection or otherwise identified by supplier. The materials manager does know that 45% of the components are purchased from $S_1$, 35% purchased from $S_2$ and the remaining from $S_3$. From past records it is also known that 6% of components purchased form $S_1$ are below standard, 8% of the components purchased from $S_2$ are below standard and 11% of the components purchased from $S_3$ are below standard. The materials manager randomly selects a component and found it below standard. From which supplier the component is most likely purchased? Draw a tree diagram.

## UNIT 3: PROBABILITY DISTRIBUTION

### Contents


3.0 Aims and Objectives

3.1 Introduction

3.2 Random variables

3.2.1 Discrete Random Variable

3.2.2 Continuous Random Variable

3.3 Discrete Probability Distribution

3.3.1 Constructing Probability Distribution

3.3.2 Mean and Advance of a Discrete Probability Distribution

3.3.3 Binomial Probability Distribution

3.3.4 Hypergeometric Probability Distribution

3.3.5 Poisson Probability Distribution

3.4 Continuous /Normal/ Probability Distribution

3.4.1 Normal Approximation to the Binomial

3.4.2 Normal Approximation to the Poisson
3.5 Answers to Check Your Progress
3.6 Model Examination Question

## 3.0 AIMS AND OBJECTIVES

In this unit, you will be introduced to repeated experiments where the result of the experiment produces two different and many possible outcomes. You will learn how to compute probabilities involving two-outcome situation using special probability formulas.

After completing this unit you will be able:

- ☞ to understand the types of random variables
- ☞ to calculate the expected value and variance of a discrete random variable
- ☞ to identity the characteristics of the binomial, hyper geometric and poison probability distributions
- ☞ to calculate probabilities for random variables following the binomial, hyper geometric and poison distributions
- ☞ to calculate the mean and variance of the binomial, hyper geometric and position distribution
- ☞ to identity the characteristics of the continuous probability distribution and its accompanying normal curve
- ☞ to calculate probabilities of a continuous random variable
- ☞ to approximate the normal distribution to the binomial and the poison distributions.

## 3.1 INTRODUCTION

Probability distribution is listing all possible values of the random variable with corresponding probabilities. The outcome of the experiment is either a success or failure. The number of ways to get certain number

of successes will determine the value that the random variable will assume.

## 3.2 RANDOM VARIABLE

Random variable is a variable whose value is determined by the out come of an experiment. That is random variable represents an uncertain outcome or it can be defined as a quantity resulting from a random experiment that by chance, can assume different values.
A random variable may be either discrete or continuous

### 3.2.1 Discrete Random Variable

Is a variable that can assume only certain clearly separated values resulting from account of some item of interest?
Example:

- The No. of employees absent in a given day
- Toss two coins and count the number of heads
- Number of defective products produced in a factory at a given shift or day or month.
- Number of customers entering to a bank in an hour time.

Is should be noted that a discrete random variable can in some cases assume fractional or decimal values. These values must be separated i.e have distance between them eg. The score of a student in a given test can be 8.5 or 7.5 such values are discrete b/se there is a distance b/n scores. There is a fixed gap between scores. You can easily list all possible values clearly and separately. If the number of students in a classroom is 35, you know the next succeeding value will be 36 there is no another value in between.

### 3.2.2 Continuous Random Variable

A variable that can assume any value in an interval. It can assume one of an infinitely large number of values. Mostly results of measurement

Example - The distance b/n two cities
- The weight of a person.
- The rate of return on investment
- The time that a customer must wait to receive his changes.

The values are not clearly separated. It is not possible to exhaustively list possible values of the random variable. If the distance between two cities is 300 km. You cannot estimate or identify the next higher distance. There are infinitely very large number of values.

## 3.3 DISCRETE PROBABILITY DISTRIBUTIONS

The values assumed by a discrete random variable depends upon the out come of an experiment. Since the out come of the experiment will be uncertain the value assumed by the random variable will also be uncertain.

The probability distribution of a discrete random variable is listing of all the outcomes of an experiment and the probabilities associated with each out come The probability distribution of a discrete random variable is a table, graph or formula that gives the probability associated with each possible value that a random variable can assume or if we organize the value of a discrete random variable in a probability distribution the distribution is called a Discrete Probability distribution. In this unit we will discuss three types of discrete probability distribution.

***Binomial, Hyper geometric and Poisson***

We denote probability distribution of a random variable x as p(x). We can sometimes use the sample space of an experiment and probability rules.

Example: Consider a test consisting of three true or false questions

The sample space consists

CCC CC| C|C |CC

C| | |C| | |C | | |

We assume;

- The student blindly guesses the answer to each question. Then each out come will be equally tickly i.e each having a probability 1/8.
- Since the student guesses blindly then the probability of answering each question correctly is ½ and the probability of answering incorrectly is also ½
- Since each question will be answered independently it follows that we can obtain the probability of each sample space out come by multiplying together the probabilities of correctly ( or incorrectly) answering individual questions.
- There fore, by independence, the probability of the samples space out come.

CCC, answering all the three questions correctly,

P(CCC) = p(c) p(c) p(c) = (½) (½) (½) = $^1/_8$

Similarly the probability of the sample space outcome CCI is

P(CCI) = (1/2 ) (1/2 ) (1/2 )= 1/8

We define the random variable X to be the number of questions that the student answers correctly. X can assume the values 0 , 1 , 2 , or 3 . Then if x = 1 one question will be answered

Correctly if and only if we obtain one of the sample space outcomes C| |

|C| | |C then

P(X=1) = P(C| |) + P(|C|) + P(| |C)

=1/8 + 1/8 + 1/8 = 3/8

Finding the probability distribution

| **Value of X = The No. of correct Answers** | **Sample space out comes correspon ding to X** | **probability of sample space out come** | **P(X) = probability of the value of X** |
|---|---|---|---|
| X=0(no correct answer) | \| \| \| | ½ x ½ x ½=1/8 | P(0) = 1/8 |
| X=1(one correct answer) | C\| \| | ½ x ½ x ½=1/8 | |
| | \|C\| | ½ x ½ x ½=1/8 | P(1)= 1/8 + 1/8 +1/8 =3/8 |
| | \| \|C | ½ x ½ x ½=1/8 | |
| X=2(two correct answers) | CC1 | ½ x ½ x ½=1/8 | |
| | C1C | ½ x ½ x ½=1/8 | P(2)= 1/8 + 1/8 +1/8 =3/8 |
| | 1CC | ½ x ½ x ½=1/8 | |
| X=3(three correct answers) | CCC | ½ x ½ x ½=1/8 | P(3) = 1/8 |

Summary: probability distribution of x

| **X, number of question Answered correctly** | **P(X) , probability of X** |
|---|---|
| 0 | P(0) = P( X=0) 1/8 |
| 1 | P(1) = P(X=1) 1/8 |
| 2 | P(2) = P(X=2) 1/8 |
| 3 | P(3)= P(X=3) 1/8 |

Sum I

**Example: 2** Suppose that the student taking the test has studied hard and does not have to guess at the answer, suppose that there is now a 90% chance that the student will answer each of the questions correctly. The probability distribution will be:

| **X** | **Sample space** | **Probability of sample space** | **P(X)** |
|---|---|---|---|
| X=0 | 111 | 0.1 X 0.1 X 0.1 = 0.001 | P(0) =0.001 |
| X=1 | C11 | 0.9 X0.1 X 0.1 =0.009 | |
| | 1C1 | 0.1 X 0.9 X 0.1 = 0.009 | P(1)=0.009+0.009+0.009 = 0.027 |
| | 11C | 0.1 X 0.1 X0.9 =0.009 | |
| X=2 | CC1 | 0.9 X 0.9 X0.1 =0.081 | P(2)=0.081 +0.081 +0.081= 0.243 |
| | C1C | 0.9 x0.1 x0.9=0.081 | |
| | 1CC | 0.1 x0.9 x0.9=0.081 | |
| X=3 | CCC | 0.9 x 0.9x 0.9 = 0.729 | P(3)=0.729 |

Similarly the distribution can be summarized Sum 1

| **X** | **P(X)** |
|---|---|
| 0 | P (0) = P (X=0) 0.001 |
| 1 | P (1) = P (X=1) 0.027 |

| | | |
|---|---|---|
| 2 | P (2) = P (X=2) | 0.243 |
| 3 | P (3) = P (X=3) | 0.729 |
| | Sum | 1 |

Properties of discrete probability distribution

1. P (X) ≥ 0 for each value of X
2. $\sum$ P (X)=1

## Check Your Progress -1

Suppose a newly married couple plans to have four children. Naturally they are curious about the sex of their children and want to estimate the outcome. Defining the event 'G' that the child will be a girl and 'B' that the child is a boy, construct the probability distribution for the number of Boys and Girls.

## 3.3.2 The Mean, Variance, and Standard Deviation of a Discrete Probability Distribution

### *3.3.2.1 Mean*

If the values of the random variable X are observed on the repetition and recorded, we would obtain the population of all possible observed values of the random variable X. This population has a mean or expected value of X.

$\mu_x$ denotes the mean of the random variable X. It is also called the expected Value of X as denoted by E(x)

$\mu_x$ = Multiply each value of X by its probability P(X) and then sum the resulting products over all possible value of X.

That is

$$\mu_x = \sum_{All\ x} x\ p(x)$$

**Example.** A car dealer has established the following probability distribution for the number of cars he expects to sell on a particular Saturday.

| Number of cars sold (X) | Probability P(x) |
|---|---|
| 0 | 0.10 |
| 1 | 0.20 |
| 2 | 0.30 |
| 3 | 0.30 |
| 4 | 0.10 |
| | Sum .1 |

On a typical Saturday ,how many cars should the dealer expect to sell?

$\mu = E(x) = \Sigma [xp(x)]$

$= 0.(0.1) + 1(0.2) + 2(0.3) + 3(0.3) + 4(0.1) = 2.1$ cars.

In the long run the dealer expects to sell 2.1 cars. On a large number of Saturdays.

**Example 2:**

Monthly sales of a certain product are believed to follow the following probability distribution. Suppose that the company has fixed monthly production cost $8,000 and that each item brings $2. Find the expected monthly profit from product sales

| **No. of items x** | **p(x)** |
|---|---|
| 5000 | 0.2 |
| 6000 | 0.3 |
| 7000 | 0.2 |
| 8000 | 0.2 |
| 9000 | 0.1 |

1

E/h(x) = $\sum_{all\ x} h(x)\,p(x)$

**Solution:**

h(x) = 2x – 8000

| x | h(x) | p(x) | h(x)p(x) |
|---|---|---|---|
| 5000 | 2000 | 0.2 | 400 |
| 6000 | 4000 | 0.3 | 1200 |
| 7000 | 6000 | 0.2 | 1200 |
| 8000 | 8000 | 0.2 | 1600 |
| 9000 | 10000 | 0.1 | 1000 |

1E[h(x)] = 5400

The expected value of a linear function of a random variable

E(ax + b) = aE(x) + b

Where a and b are fixed numbers once we know the expected value of x, the expected value of ax + b is just aE(x) + b.

In the above example we could have obtained the expected profit by finding the mean of x first and then multiplying the mean of x by 2 and subtracting from this the fixed cost of 8000.

The mean χ is 6, 700 and the expected profit is therefore E[h(x)] =
E(2x – 8000) = 2E(x) – 8000 = 2(67,000) – 8000 = 5400

***3.3.2.2 Variance and Standard Deviation of the Discrete Probability Distribution***

The mean does not describe the amount of spread or variation of a distribution. The variance and standard deviation allows us to compare the variation in two distributions having the same mean but different spread.

The formula for the variance of a discrete probability distribution is

$\delta^2 = \Sigma[(x - \mu)^2\ p(x)]$ $\qquad\qquad$ $\delta = \sqrt{\sum (x - \mu)^2\ p(x)}$

or

$E(x^2) - [E(x)]^2$ where

$Ex^2$ = the expected value of $x^2$ i.e., $\Sigma x^2\ p(x)$

E(x) = the expected value of x

**Example.** For the car dealer find the variance and standard deviation

| X | p(x) | (x - μ) | $(x - \mu)^2$ | $(x - \mu)^2\ p(x)$ |
|---|---|---|---|---|
| 0 | 0.1 | 0 – 2.10 | 4.41 | 0.441 |
| 1 | 0.2 | 1 – 2-10 | 1.21 | 0.224 |
| 2 | 0.3 | 2 – 2.10 | 0.01 | 0.003 |
| 3 | 0.3 | 3 – 2.10 | 0.81 | 0.243 |
| 4 | 0.1 | 4 – 2.1 | 3.61 | 0.361 |
| | 1 | | | $\delta^2 = 1.29$ |

$\delta^2 = 1.29$

$\delta = \sqrt{1.29}$

= 1.136 cars

Using the other formula we will have the same variance and standard deviation

| X | p(x) | $x^2$ | x p(x) | $x^2\ p(x)$ |
|---|---|---|---|---|
| 0 | 0.10 | 0 | 0 | 0 |
| 1 | 0.20 | 1 | 0.2 | 0.2 |
| 2 | 0.30 | 4 | 0.6 | 1.20 |
| 3 | 0.30 | 9 | 0.9 | 1.70 |
| 4 | 0.10 | 16 | 0.4 | 1.60 |
| | | | $\mu = 2.1$ | $Ex^2 = 5.7$ |

$\delta^2 = E(x^2) - [E(x)]^2$

$= 5.7 - (2.1)^2$

$= 5.7 - 4.41$

$= 1.29$

$\delta = \sqrt{1.29} = 1.136$

## Check Your Progress –2

Find the variance and standard deviation of the distribution of correct answer answered by the student with 0.90 probability of answering each of the three questions correctly.

### 3.3.3 The Binomial Distribution

The binomial distribution is a discrete probability distribution

The binomial distribution has the following characteristics.

1. The experiment consists of N identical trials and the data collected are the results of counts.
2. An out come of an experiment is classified into one of two mutually exclucle categories a success or failure. i.e each trial results in a success or failure.
3. The probability of success remains the same for each trial. So does the probability of a failure. This implies that the probability of failure of any trial is 1- (probability of success). Probability of success is denoted by p and probability of failure by q of then q = 1 - p
4. The trials are independent i.e the outcome of one trial does not affect the outcome of any other trial.

**Example 1**. Suppose that 40% of all customers who enter a department store make a purchase.

What is the probability that 2 of the next 3 customers will make a purchase?

Note that this problem qualifies all the characteristics of the binomial distribution

- The trials are three and each of the three customers will either purchase or not purchase so the three trials are identical
- The outcome of each trial will result in either a purchase (success) or not purchase (failure)

- The probability of purchase is the same 0.4 for each of the three customers. And probability of failure (not purchase) will be 1 – 0.4 = 0.6 for each.
- The decision of one customer will not affect the decision of others. I.e., decision to purchase or not to purchase by each customer is independent.

The sample space of this trial consist of eight-sample space out comes.

SSS SSF SFS FSS

FFS FSF SFF FFF

S is a success (purchase)

F is a failure (not purchase)

Two out of three customers make a purchase if one of the sample space out come SSF, SFS, FSS occurs. By independent

$P(SSF)= P(S)\ P\ (S)\ P(F) = (4)\ (4)\ (.6) = (0.4)^2\ (0.6)$

$P(SFS)= P(S)\ P(F)\ P(S) = (0.4)\ (0.6)\ (0.4) = (0.4)^2\ (0.6)$

$P(SSF)= P(F)\ P\ (S)\ P(S) = (0.6)\ (0.4)\ (0.4) = (0.4)^2\ (0.6)$

Then the probability that two out of the three customers make a purchase is

$P(SSF) + P\ (SFS) + P\ (FSS)$

$= (0.4)^2\ (0.6) + (0.4)^2(0.6) +(0.4)^2(0.6)$

$= 3(0.4)^2(0.6)$

Note that:

1. The 3 is the number of sample space out come (SSF, SFS and FSS) that correspond to the event i.e., two out of the three customers make a purchase. This equals the number of ways we can arrange two successes among three trials.
2. 0.4 is P, the probability that a customer makes a purchase
3. 0.6 is q = 1 – P , the probability that a customer does not make a purchase.

Therefore, the probability that two of the next three customers make a purchase is

= (the number of ways to arrange 2 success among 3 trials) $P^2q^1$

Notice that SSF, SFS, FSS each of these sample space out comes consists of two successes and one failure. The probability of each of these sample space out comes equals $(0.4)^2(0.6)^1 = p^2q^1$

P is raised to a power that equals the number of successes (2) in the three trials and q is raised to a power of failures (1) in the three trials.

In general, each of the sample space out comes describing the occurrence of X successes (purchase) in n trials represents a different arrangements of X success in n trials. However each of these sample space outcomes consist of X successes and n – X failures. There fore, the probability of each sample space outcome is

$P^xq^{n-x}$ it follows by analogy that the probability that X of the next n trials are successes (purchase) is

(The number of ways to arrange X successes among n trials) $(P^xq^{n-x})$
The number of ways to arrange X successes among n trials equals.

$\frac{n!}{X!(n-x)!}$ n! is read n factorial

n! = n(n – 1) (n – 2) … (n – n)!

(n – n) = 0; 0! = 1 by definition

Then we call x a binomial random variable and the probability of obtaining X success in n trials is

P (X) = $\left(\frac{n!}{X!(n-x)!}\right)(p^xq^{n-x})$ ⇒ The Binomial formula

For the above example we can solve for p(x = 2) as follows

n = 3

p = 0.4

q = 0.6

$$p(x = 2) = \frac{3!}{2!(3-1)!} 0.4^2\ 0.6^1$$

$$= \underline{\underline{0.288}}$$

**Example 2:** An examination consists of four true or false question and student has no knowledge of the subject matter. The chance that the student will guess the correct answer to the first question is 0.5. a) What is the probability of getting exactly none out of four correct?

N = 4 p = 0.5 q = 0.5 x = 0

$$P(X) = \frac{n!}{x!(n-x)!} P^x q^{n-x}$$

$$P(X=0) = \frac{4!}{0!(4-0)!} 0.5^0\ 0.5^4 = 0.0625$$

b) What is the probability of getting exactly one out of four correct

$$P(1) = \frac{4!}{1!(4-1)!} (0.5^1)\ (1-0.5)^{4-1} = 0.2500$$

The probability of getting exactly 0, 1 , 2 , 3 or 4 correct out of a total of four questions is shown in the table for the Binomial probability distribution.

| Number of correct guess (x) | Probability (x) |
|---|---|
| 0 | 1/6 = 0.625 |
| 1 | 4/16=0.2500 |
| 2 | 6/16=0.3750 |
| 3 | 4/16 = 0.2500 |
| 4 | 1/16= 0.0625 |

| | |
|---|---|
| Total | 16/16 = 1 |

**Check Your Progress –3**

A truck operator has determined that a car repair shop derivers maintained trucks on schedule 60% of the time. If the operator has 6 trucks under maintenance (a) construct the probability distribution for the number of truck to be delivered on time. (b) Find the expected value and the standard deviation of the distribution.

**Using the Binomial Probability Table:**

A binomial probability distribution is a theoretical distribution, can be generated mathematically. However except for problems involving small n the calculations for the probabilities of 0, 1, 2 success can be rather tedious. As an aid in finding the needed probabilities of 0,1,2,3 . . . successes for various values of n and P an extensive table has been developed.

The table has up to n =25 or 30

P from 0.05 .0.1,0.2. . . 0.90 , 0.99

X= from 0-25 or 30

**Example.** 25% of college students in a classroom join the HIV AIDS prevention club. If 20 students are enrolled in the class, what is the probability that two or fewer will join the club?

**Solution:**

P = 0.25

n = 20 then

$p(x \leq 2) = p(0) + p(1) + p(2)$ from the table

p(0) = 0.0032

p(1) = 0.0211

p(2) = 0.0660

Sum = p(x ≤ 2) = 0.0909

In similar fashion you can find the probability for any value of x using the table.

**The mean, variance and standard deviation of a Binomial Random Variable**

If X is a binomial Random variable then $\mu_x$ = np the mean $\mu$, of the distribution is equal to np

The mean is equal to the number of trials n, times the probability of success in a single trial, p.

**Example 1.** The number of heads appearing in five tosses of a fair coin.

E(x) = n p = 5(0.5) =2.5

As a long run average, we expect that 2.5 out of 5 tosses of a fair coin will result in heads.

The variance of a binomial X is, $\sigma^2$ and $\sigma^2$ = npq

The standard deviation is $\sigma = \sqrt{\sigma^2} = \sqrt{npq}$

**Example 2**. 35% of the students registered in the 1st semester join the marketing department. If 1000 students are registered,

(a) How many of them are expected to join the marketing department

$\mu$ = np

$\mu$ - 1000 (0.35) = 560

(b) What is the standard deviation?

$\delta = \sqrt{npq}$

$= \sqrt{1600(0.35)(0.65)}$

$= \sqrt{364}$

### 3.3.4 Hyper Geometric Distribution

The binomial distribution is appropriate when we are sampling from a population that is much larger than the sample. The Binomial assumes sampling with replacement.

We sample an item, whether it is a success or failure, returne or put it back to the population before the next item is selected for the sample, then we are sampling with replacement.

Sampling with replacement is not a frequently used procedure and most sampling is done without replacement. Thus the outcomes are not independent and the probability for each successive observation or trial will change.

Since the probability of success, does not remain the same from trial to trial the binomial distribution should not be used.

**Example.** If you draw a card (without replacement) from a standard deck of 52 playing cards what is the probability of getting the first card a king and the second a queen? P($1^{st}$k n $2^{nd}$Q)

= p(k) p(Q/ $1^{st}$k) = $\frac{4}{52}\left(\frac{4}{51}\right) = \left(\frac{16}{2652}\right)$

Note that probability of success for the $1^{st}$ card was $\frac{4}{52}$ while for the $2^{nd}$ card $\frac{4}{51}$ i.e., probability of success changes.

If a sample is selected from a small population with out replacement the hyper geometric distribution should be applied.

Since we sample from a large population the hyper geometric distribution is less use full than the binomial.

***Derivation of the hyper geometric distribution***

Consider a collection of N objects which S of these objects have a certain attribute and the remaining N – S objects do not have this attribute. If a sample of n objects is chosen at random and with out replacement from this collection of objects, then the number of objects in the sample having the attribute is a random variable having a hyper geometric distribution.

To find the probability distribution for X we follow the following arguments.
Since the n objects are chosen randomly from the N objects available, there are

${}_{N}C_{n}$ different possible subset of n objects that could be chosen. To find p(x) we need to know the number of these subsets that have X objects having the attribute ( and n – x objects not having the attribute) . There are ${}_{S}C_{X}$ ways of choosing X objects from the S having the attribute in the population.

${}_{N-S}C_{n-x}$ ways of choosing n – x objects from the N-S not having the attribute. The quantities n, N and S and parameters of this distribution as indicated by the following notation

$$P(X) = \frac{S^{C}X \; ( N\text{-}s \; C \; n - X )}{N^{C}n}$$

Where:

N: the size of the population

S- the number of success (objects with certain attributes) is the population

X- the number of success (of interest) objects in the sample having the attribute n is the size of the sample (objects chosen randomly from the population)

**Example 1**. An inspector is to examine a population of 20 shipping orders to check for authorized credit approval. If 15 of these have authorized credit approval and if a sample of 4 orders is to be randomly chosen,

what is the probability that exactly 3 will have authorized credit approval?

Since the orders are chosen, at random, we know that all subsets of 4 orders from the 20 are equally likely to be chosen. By using the equally likely outcomes approach, we see that there are

$$20\ C\ 4 = \frac{20!}{4!(20-4)!} = \frac{20!}{4!\,(16!)} = 4845$$

Ways that a sample of four can be chosen out of 20.

There are $15C_3 = 455$ ways that three credit approved orders can be selected from 15 credit approved orders and

${}_sC_1 = 5$ ways that one non –approved order can be selected from five non-approved order consequently.

$$P(x=3) = \frac{(15C3)\ (\ 5C_1)}{20c4} = \frac{455\ (5)}{4845} = \underline{0.4696}$$

**Example 2.** Suppose that automobiles arrive at a dealer's shop in lots of 10 and that for time and resource considerations only 5 out of each 10 are inspected for safety. The 5 cars are randomly chosen from the 10 in the lot.

If 2 out of the 10 cars in the lot are bellow standards for safety, what is the probability that at least 1 out of the 5 cars to be inspected will be found not meeting the safety standard?

N = 10

S = 2

N = 5

X = at least one i.e., one o

$$P(x=1) = \frac{2C_1(10_{-2}C_{5-1})}{10C_5} = 0.556$$

$$p(x = 2) = \frac{2c_2(_{10-2}C_{5.2})}{10C_5} = 0.222$$

p(at least one) = p(1) + p(2)

= 0.556 + 0.222 = 0.778

## Check Your Progress –4

Suppose 50 TV sets were manufactured during the week. 40 operated perfectly and 10 had at least one defect.

A sample of 5 is selected at random. What is the probability that 4 of the 5 will operate perfectly?

**Mean and Variance of the Hyper Geometric Distribution**

If X is a random variable having a hyper geometric distribution with parameters n, N and S then. E(x) = $n\left(\frac{S}{N}\right)$ and $\delta_x^2 = n\left(\frac{S}{N}\right)\left(1-\frac{S}{N}\right)\left(\frac{N-n}{N-1}\right)$

**Example.** If 180 out of 200 shipping orders that the inspector will examine have authorized credit approval what are the mean and variance of the number in a sample of 40 randomly chosen orders that will have credit approvals?

E(x) = 40 (180/200) =36

$\delta^2 x$ = 4(180/200) (20/200) (160/199)=2.8945

## 3.3.5 The Poison Probability Distribution

The third important discrete probability distribution is the Poisson. The Poisson distribution counts the number of successes in a fixed interval of time or with in specified regions.

Eg. The number of machine failure in a week

- the number of traffic accidents per month in town
- the number of emergency patients arriving at a hospital in an hour
- the number of orders received per day
- the number of defects in a square metere metal sheet.

To apply the Poisson distribution the following condition are required

1. The probability of success in a short interval of time (or space) is proportional to the size of the interval. If we count 6 patients arriving in an hour then we expect 3 in half an hour and 2 in 20 minutes.
2. In a very small interval, the probability of successes is close to zero. If 6 patients arrive in an hour we expect none in 10 seconds.
3. The probability of success in a given interval is independent of where the interval begins.
4. The probability of success over a given interval is independent of the number of the events that occurred prior to the interval.

The Poisson distribution is described mathematically by the formula.

$$P(x) = \frac{\mu^x e^{-\mu}}{X!}$$

Where;

$\mu$ is the mean number of success /average rate/

e is the base of natural logarithm or mathematical constant with value 2.7183

X is the number of success in the interval

P (X) is the probability of X successes in an interval

The Poisson distribution can be used to approximate the binomial distribution when the probability of a success is small and the number of trial is very large.

Usually the probability of success become quite small after few occurrences as the random variable X for a Poisson distribution assume an infinite number of values.

**Example1.** Assume that billing clerks rarely make errors in data entry on the billing statements of a co. Many statements have no mistakes; some

have one, a very few have tow mistakes; rarely will a statement have three mistakes; and soon. A random sample of 1000 statements revealed 300 errors. What is the probability of no mistakes appearing in a statement $\mu = 300/1000=0.3$

$$P(0) = \frac{0.3^0(2.7183)^{-0.3}}{0!} = 0.7408$$

**Example 2.** A bank manger wants to provided prompt service for customers at the banks drive up window. The bank currently can serve up to10 customers per 15-minute period with out significant delay. The average arrival rate is 7 customers per 15minute period. A assuming X has a Poisson distribution find the probability that 10 customers. Will arrive in a particular 15-minute period.

$\mu = 7$

X= 10

$$P(10) = \frac{7^{10}\ 2.7183^{-7}}{10!} = 0.710$$

## Check Your Progress –5

A telephone company's goal is not to have more than there line failures in a particular 1km line. Currently the company is experiencing four line failures in 1km line.

a) what is the probability that the company will meet its goal?
b) what is the probability that the company will not meet its goal?

**Variance and standard deviation of the passion probability distribution**

The variance of the poison distribution is equal to the mean of the distribution.

$\delta^2 = \mu$ then

$\delta = \sqrt{\mu}$

## 3.4 THE NORMAL / CONTINIOUS / PROBABILITY DISTRIBUTION

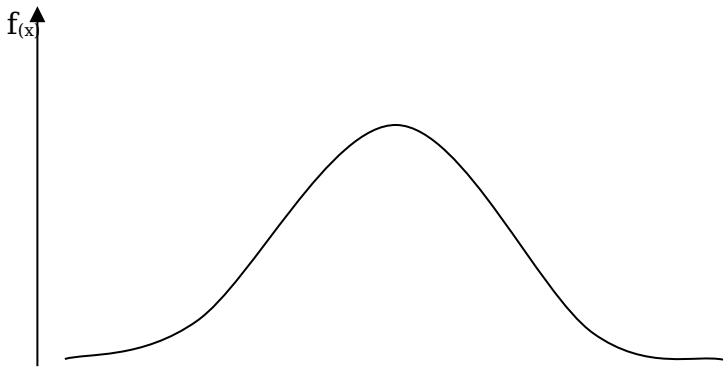
As noted earlier in this unit a continuous random variable is one that can assume an infinite number of possible values with in a specified range. It usually results from measuring some thing.

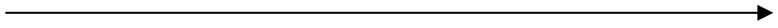
It is not possible to list every possible value of the continuous random variable along with a corresponding probability.

The most convenient approach is to construct a probability curve. The proportion of area included between any two point under the probability curve identified the probability that a randomly selected continuous variable has a value between those points.

Characteristics of a normal probability distribution and its accompanying normal curve

1. The normal curve is bell – shaped and has a single peak at the exact center of the distribution. The arithmetic mean median and the mode are equal and are located at peak. Thus half the area under the curve is above this center point, and the other half is below it.
2. The normal probability distribution is symmetrical about its mean. If we cut the normal curve vertically at this central value, the two halves will be mirror images.
3. The normal curve falls of smoothly in either direction from the central value. It is asymptotic, meaning that the curve gets closer and closer to the X – axis but never actually touches it. In real world problems, however, this is somewhat unrealistic. The






***The Normal Curve***

The normal probability distribution is important in statistical inference for three distinct reasons:

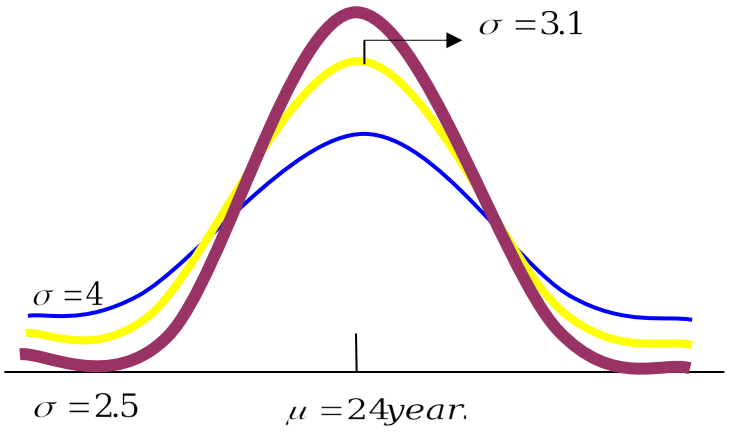
1. The measurements produced in many random processes are known to follow this distribution.
2. Normal probability can often be used to approximate other probability distribution, such as the binomial and Poisson distributions.
3. Distribution of such statistics as the sample mean and sample proportion often follow the normal distribution regardless of the distribution of the population.

**Constructing the Probability Curve**

There is not just one normal probability distribution. There is a family of them we night have one of the following:

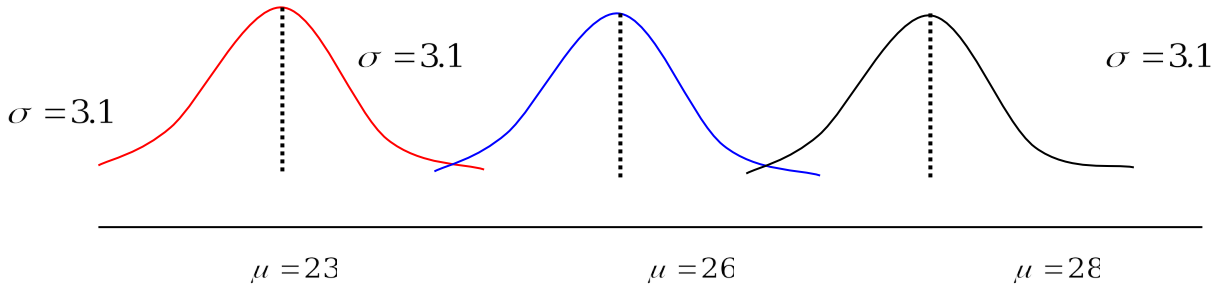
a. Equal means and different standard deviations

eg. Average age of students in three sections S1, S2, S3 is equal 24 years. But the standard deviation for S1 =2.5, S2 = 3.1 and S3 = 4.



The shape of the curves is determined by the standard deviation. The smaller the standard deviation the more packed the curve will be and the larger the standard deviation the more flat and wider the curve will be

b. different means but equal standard deviation. Both sections have equal standard deviation 3.1 but different means S1=23 S2=26 S3=28

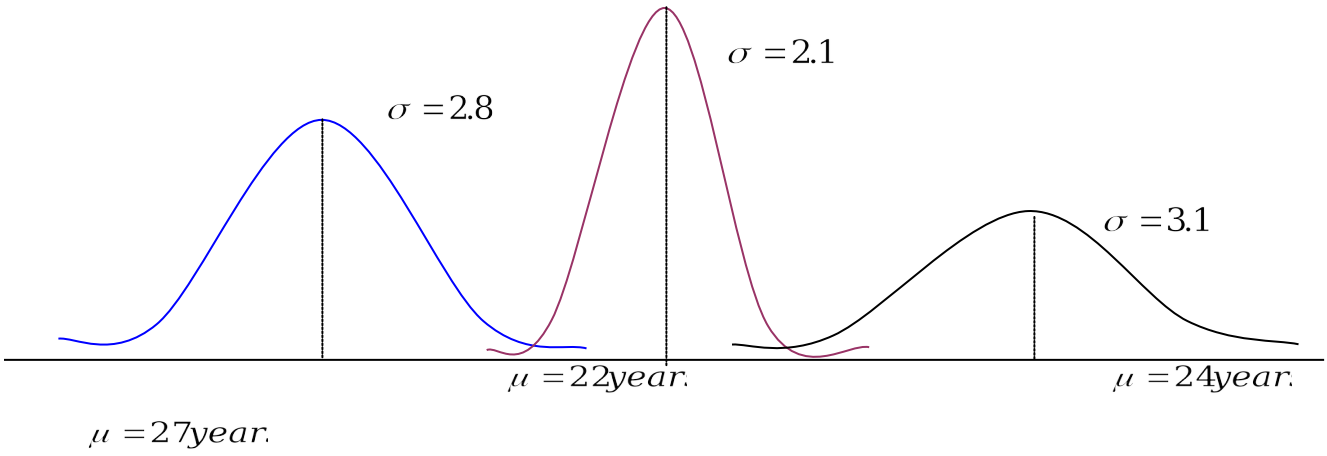



c. Different means and different standard deviations

For S1 μ = 22 and δ =2.8

S2 μ =24 and δ =2.1

S3 μ =27 and δ =3.1



The number of normal distributions is unlimited. It would be practically impossible to provide a table of probabilities (as the binomial and Poisson) for each combination of $\mu$ and $\delta$ or using a formula.

One member of the families of normal distributions can be used for all problems where the normal distribution is applicable.

It has a mean of 0 and a standard deviation of 1 and is called Standard Normal Distribution.
First it is necessary to convert or standardize the actual distribution to a standard normal distribution using Z value. Z is called the normal deviate.

Z value is the distance between a selected value and the population mean in units of the standard deciation.

**Transformation of the Normal Random Variable**
Since there are infitely many possible normal random variables one of them is selected to serve as our standard.

We want to transform X in to the standard normal random variable Z.

**Example.** We have a normal random variable X with $\mu$ =50 and $\sigma$ =10 we want to convert this random variable with $\mu$ =0 and $\sigma$ =1.

We move the distribution from its center of 50 to a center of 0. this is done by subtracting 50 from all the values of X. Thus we shift the distribution 50 units back so that its new center is 0. If we subtract the mean from all values of X, the new distribution (X-$\mu$) will have a mean of zero.

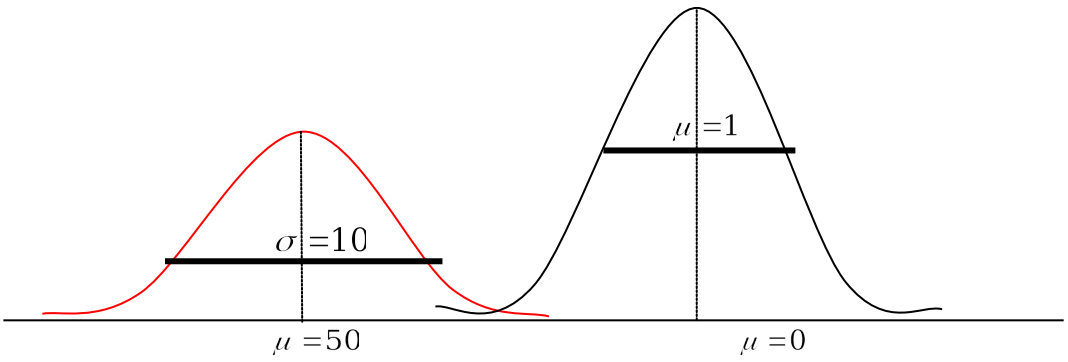
The second thing we need to do is to make the width of the distribution, standard deviation equal to 1. This is done by squeezing the width of the distribution down from 10 to 1. Because the total probability under the

curve must remain 1. the distribution must grow up ward to maintain the same area.

Mathematically, squeezing the curve to make the width 1 is equivalent to dividing the random variable by its standard deviation. The area under the curve adjusted so that the total remains the same.

The mathematical transformation from X to Z is thus achieved by first subtracting μ from X and then dividing the result by σ.

$$Z = \frac{X - \mu}{\sigma}$$




**Example** – The weekly incomes of a large group of middle managers are normally distributed with a mean of 1000 Br. and standard deviation of Br. 100. What is the Z value for an income of

a) Br. 1100? $Z = \frac{X - \mu}{\sigma}$ $\mu = 1000$ $\sigma = 100$

$$Z = \frac{1100 - 1000}{100} = 1$$

This means an income of 1100 is one standard deviation above the mean.

b) Br 900?

$$Z = \frac{900 - 1000}{100} = -1$$

This implies that an income of Br. 900 is one standard deviation (Br. 100) below the mean.

c) Br. 1250?

$$Z = \frac{1250 - 1000}{} = 2.5$$

100

This implies that an income of Br. 1250 is 2.5 standard deviations above the mean

d) Br. 850?

$$Z = \frac{850 - 1000}{100} = -1.5$$

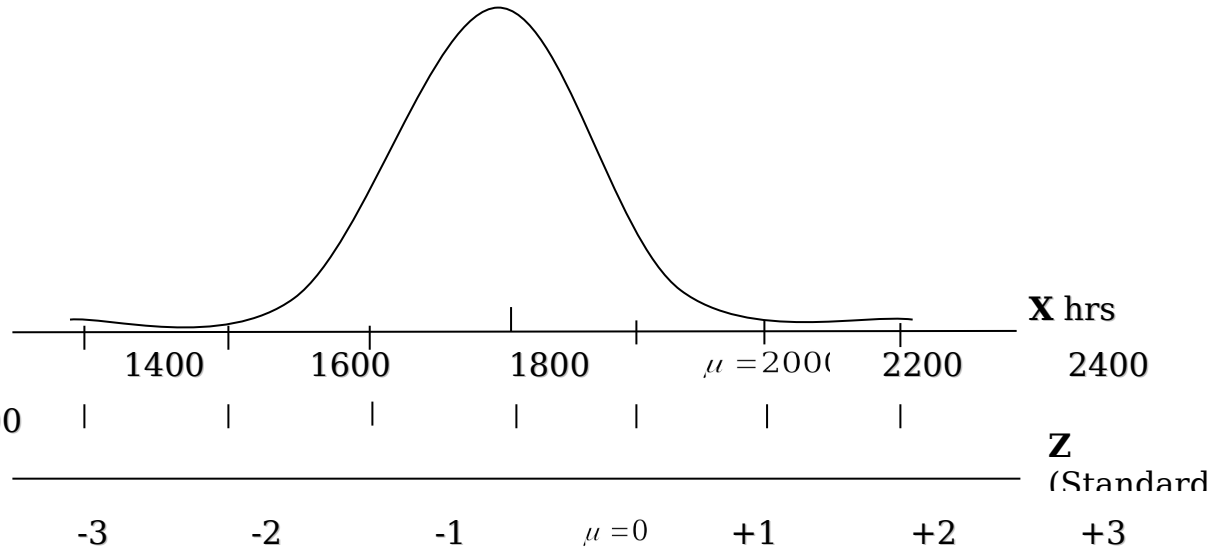
This means an income of Br. 850 is 1.5 standard deviations below the mean

**Finding probabilities using the normal probability table**

For any value of Z calculated the corresponding probability can be easily found from the Z table.

**Example 1:** The lifetime of an electrical component is known to follow normal distribution with mean 2000 hr and standard deviation 200 hr

(a) What is the probability that a randomly selected component will last between 2000 and 2400 hr?




The lower boundary of the interval is at the mean of the distribution and therefore at Z = 0. The upper boundary of the interval in terms of Z is

$$Z = \frac{\chi - \mu}{\sigma} = \frac{2400\ 2000}{200} = 2$$

By reference to the probability table

$p(0 \leq Z \leq +2) = 0.4772$

$p(2000 \leq x \leq 2400) = 0.4772$

This means a randomly selected component will have a probability of 0.4772 to last between 2000 to 2400 hr. Or we can say 47.72% of all components will last between 2000 to 2400 hr.

(b) What is the probability that a randomly selected component will last more than 2200 hrs?

Note that the total area to the right of the mean 2000 is 0.5. Therefore if we determine the proportion between the mean and 2200, we can subtract this value from 0.50 to obtain the probability of the hrs x being greater than 2200.

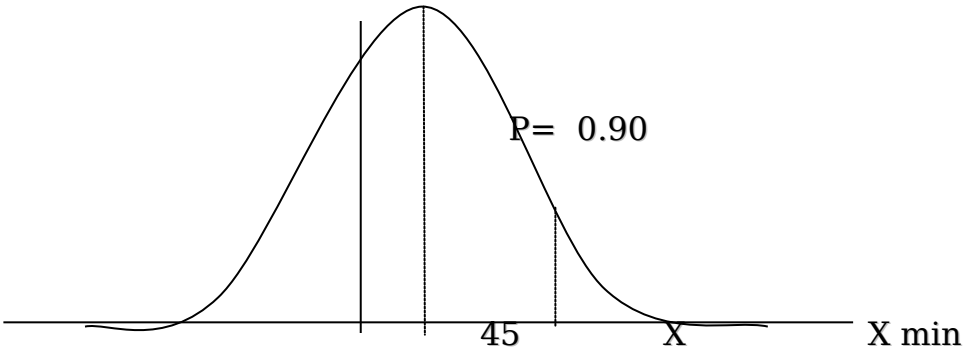
$$Z = \frac{2200 - 2000}{200} = 1$$

$p(0 \leq Z \leq +1.0) = 0.3413$

$p(Z > +1) = 0.5000 - 0.3413$

$= 0.1587$

This means 15.87% of the components will last more than 2200 hrs.




**Example 2:** The amount of time required for a certain type of car repair at a service guarage is normally distributed with the $\mu = 45$ min. And the standard deviation $\sigma = 8$ min. The service manage plans to have work begin on a customers car 10 min after the car is dropped off and he tells the customer that the car will be ready with in 1 hrs total time.

a) What is the probability that he will be wrong?

P(error) = p ( x > 50 min) , since actually work is to begin in 10 min, the actual repair must be completed in the remaining 50 min. And the manager will be wrong if the repair takes more than 50 minutes.

$Z = \frac{X - \mu}{\sigma} = \frac{50-40}{8} = +0.62$ p(Z = 0.62) = 0.2324 then,

P( x > 50) = P (Z > + 0.62 )= 0.5000 –0.2324 = 0.2676

b) What is the required working time allotment such that there is a 90%chance that the repain will be completed with in that time?

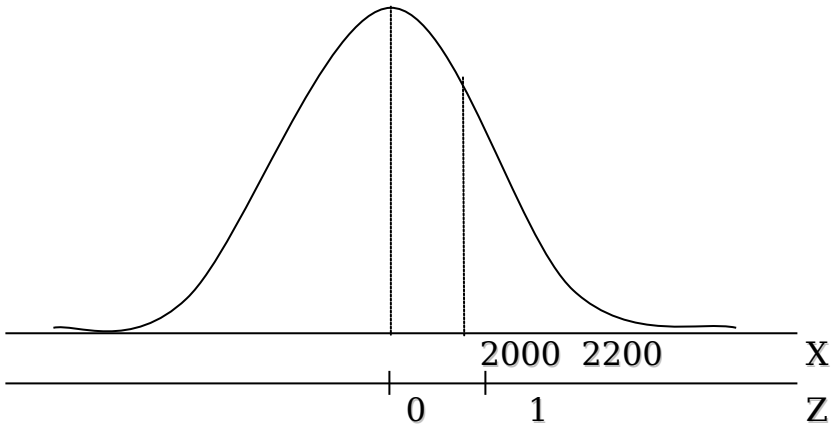
If the proportion of the area is 0.90, then because a proportion of 0.5 is to the left of the mean, it follows that a proportion of 0.4 is between the mean and the unknown value of X.

By looking the table the closest we can come to a proportion of 0.40 is 0.3997 and the Z value associated with this proportion is Z = + 1.28

Now convert Z value to a value of X

$Z = \frac{X - \mu}{\sigma}$ , $Z(\sigma) = x - \mu$, $x = \mu + Z\sigma$

X = 45 + (+1.28) (8.00)= 45 +10.24=55.24 min




This means if the service manager allots 55.24 minutes for the repair he will have a 90% chance to complete the repair with in 55.24 minutes.

C) What is the working time allotment such that there is a probability of just 30% that the repair can be completed with in that time?

Since a proportion of area of 0.3 is to the left of the unknown value of X it follows that a portion of 0.20 is between the unknown value and the mean. By reference to the table the proportion of area closest to this is 0.1985 and the Z value corresponding to this probability is 0.52. The Z value is negative because the unknown value is to the left of the mean.

$X = \mu + 2\sigma$

X = 45 + (-0.52)(8) = 40.84 min. The service manager will have a 30% chance to complete the repair with in 40.84 min.

**Example 3.** Returning again to the weekly incomes illustration, $\mu$ = 1000 and $\sigma$ =100

(a) What percent of the executive earn weekly incomes of 1245 or more?

$X \geq 1245$

$$Z = \frac{1245 - 1000}{100} = 2.45$$

The area associated with Z = 2.4 is 0.4929. This is the probability between 1000 and 1245. The probability for 1245 and beyond is found by subtracting 0.4929 from 0.5. This is equal to = 0.0075. That only 0.71% of the executives earn weekly incomes of 1245 or more.

(b) What is the probability of selecting an income between 840 and 1200

This problem is divided in to two parts

1) for the probability between 840 and the mean

$$Z = \frac{840 - 1000}{100} = -1.60$$

2) for the probability between the mean 1000 and 1200

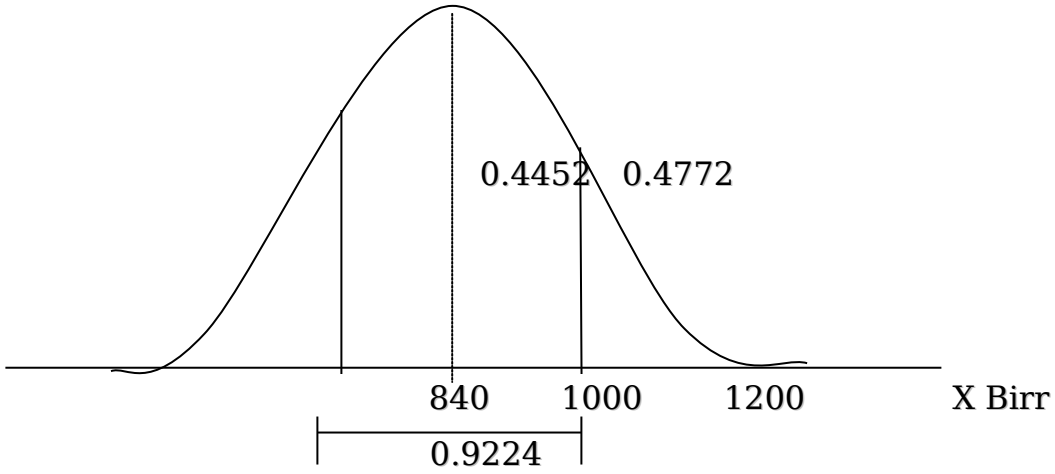
$$Z = \frac{1200 - 1000}{100} = 2$$

The probability of Z = -1.60 is 0.4452

The probability of Z = 2 is 0.4772

0.4452 + 0.4772 = 0.9224 or 92.24% i.e.,

92.24% of the managers have weekly incomes between 840 and 1200.




c) What is the probability that a randomly selected middle manager will have an income between 1150 and 1250

This problem is separated in two parts. First find the Z value associated with 1250

$$Z = \frac{1250 - 100}{100} = 2.5$$

Next find the Z value for 1150

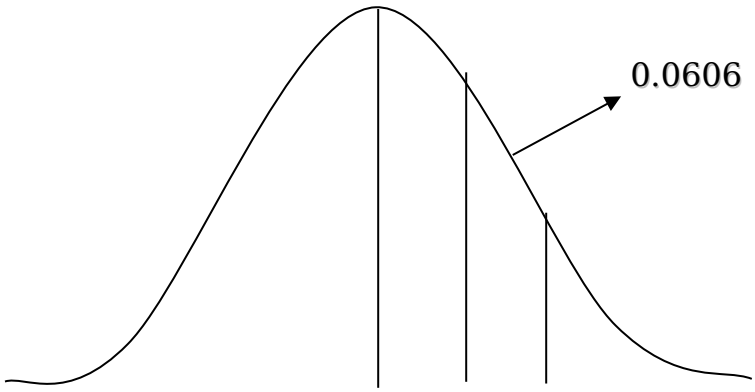
$$Z = \frac{1150 - 1000}{100} = 1.5$$

p(Z = 2.5) = 0.4935

Similarly p(Z = 1.5) = 0.4332

So the probability between 1150 and 1250 equals

0.4938 – 0.4332 = 0.0606



1000 1500 1250

## Check Your Progress –6

Service life of truck tires for heavy-duty trucks follows the normal distribution with mean 50000 km and standard deviation 5000 km.

a) What is the probability that a tyre will last between 47,000 km and 60000 km?
b) What percentage of the tyres will last below 48500 km?
c) If the supplier of the tyres is planning to replace only 1% of those tyres with the minimum performance what should be the service life for warranty?

## Computing unknown Mean and unknown Standard deviation

Some times the mean and the standard deviation of normal probability distribution may not be given or known. In such situations the probability of two unknown variables ($x_1$ and $x_2$) is used to compute the mean and standard deviation.

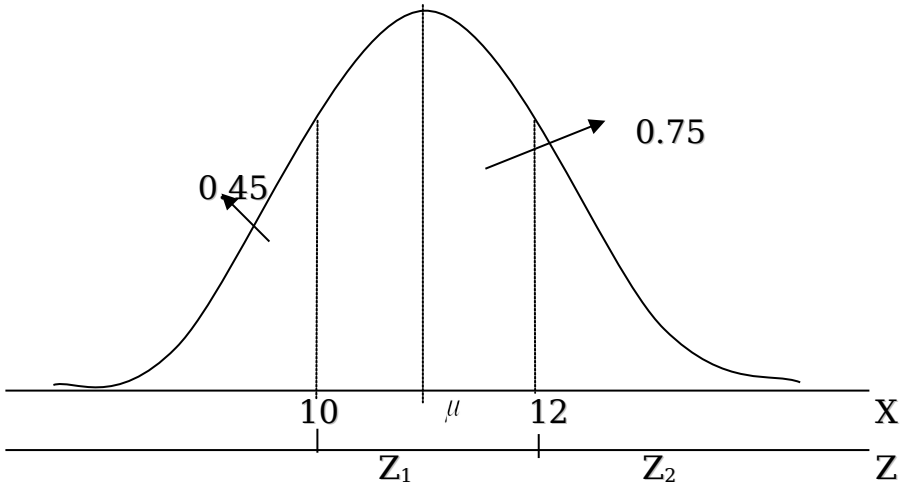
**Example 1:** The construction time for a certain building is normally distributed with an unknown mean and unknown variance. We do know, however, that 75% of the time construction takes less than 12 months and 45% of the time construction takes less than 12 months and 45% of the time construction takes less than deviation of the construction time.

We have $p(x < 12) = 0.75$ and

$p(x < 10) = 0.45$, this follows that

$p\left(Z_1 < \frac{12-\mu}{\delta}\right) = 0.75$ and

$p\left(Z_2 < \frac{10-\mu}{\delta}\right) = 0.45$




From the table we find that $Z_1 = -0.12$ and $Z_2 = 0.67$ substituting these two values for μ and σ we get: $\frac{10-\mu}{\sigma} = -0.12$ and

$\frac{10-\mu}{\sigma} = 0.67$

by cross multiplication,

$-0.12\sigma = 10 - \mu$

$0.67\sigma = 12 - \mu$

$\mu = 10 + 0.1\sigma$

$\mu = 12 - 0.67\sigma$

We have two equation with two unknown and it follows that

$10 + 0.12\sigma = 12 - 0.67\sigma$

$0.79\sigma = 2$

$\sigma = 279 = 2.53$

$\mu = 10 + 0.12\ (2.53)$

$= \underline{\underline{10.30}}$

## Check Your Progress –7

A machine is to be designed so that only 2.5% of the length of bolts made are more than 0.01 mm above the mean and only 2.5% are more than 0.01 below the mean. What standard deviation must the machine have to meet these objectives?

### Normal Approximation

One of the reasons why we apply the normal probability distribution is that it is more efficient than the binomial or poisson when these distributions involve larger n or $\mu$ values respectively.

### 3.4.1 The Normal approximation to the Binomial

The table of the binomial probabilities goes successively from an n of 1 to n of 25 or 30. Suppose a problem involved taking a sample of 60. Generating a binomial distribution for that large a number using the formula would be very time consuming. A more efficient approach is to apply the normal approximation. This seems reasonable because as n increases, a binomial distribution gets closer and closer to a normal distribution.

The normal probability distribution is generally deemed a good approximation to the binomial probability distribution when np and nq are both greater than 5.

Since there is no area under the normal curve at a single point, we assign interval on the real line to the discrete value of X by making what we call a *continuity correction factor*. Continuity correction factor is subtracting or adding, depending on the problem, the value 0.5 to a selected value when a binomial probability distribution is being approximated by a normal distribution. We add 0.5 to x when $x \leq$ and $x >$ a certain value we subtract 0.5 from x when $x <$ and $\geq$ a certain value.

**Example1:** supposes that the management of a restaurant found that 70% of their new customers return for another meal. For a week in which 80 new (first time) customers dined at the restaurant, what is the probability that 60 or more will return for another meal?
Notice that the binomial conditions are met.

To calculate this probability using the binomial formula means computing the probabilities of 60 , 61 , 62 ….. 80 and adding them to arrive at probability of 60 or more. This is quick ward the practically impossible. So the most appropriate solution is the normal approximation.
Step 1: compute the arithmetic mean and the standard deviation of the binomial distribution
$\mu$ = np = 80 (0.70) = 56
$\sigma = \sqrt{npq} = \sqrt{80(0.7)(0.3)}$ = 4.0988

Step 2. Apply continuity correction factor for x. x = 60 for the discrete random variable
60 or more means 60 inclusive. Since the lower limit for 60 is 59.5, Sixty starts from 59.5. This is similar to rounding number between 59.5 and 60.5 to 60. 60 is a value b/n 59.5 and 60.5

Step 3: Determine the standard normal value, Z,

$$Z = \frac{X - N}{\sigma} = \frac{59.5 - 56}{4.0998} = 0.85$$

Step 4: calculate the probability of a Z value greater then 0.85
The probability of Z value between O and 0.85 is 0.3023
To determine the probability of a Z value greater than 0.85 subtract 0.3023 from 0.50
0.5000 – 0.3023 = 0.1977. So the probability that 60 or more customers will come again is 19.77%

**Example 2:** For a large group of sales prospects it is known that 20% of those contacted personally by a sales representative will make a

purchase. If a sales representative contacts 30 prospects, what is the probability that 10 or more will make a purchase?

$\mu$ = np = (30) (0.2) = 6.00

$\sigma = \sqrt{npq} = \sqrt{(30)(0.2)(0.8)}$ = 2.19

10 or more is assumed to begin at 9.5. i.e., x = 9.5

$Z = \frac{9.5 - 6.00}{2.19} = \frac{3.5}{2.19} = + 1.60$

The probability for Z = 1.60 = 0.4452

p(≥1.6) = 0.5000 – 0.4452 =0.0548

### 3.4.2 Normal Approximation of Poisson Distribution

When the mean of a Poisson distribution is relatively large, the normal probability distribution can be used to approximate the Poisson distribution. For a good normal approximation to the poisson $\mu$ must be greater than or equal to 10.

**Example:** The average number of calls for a service received by a machine repair shop per 8 hr shift is 10.00. What is the probability that more than 15 calls will be received during a randomly selected 8 hr shift

$\mu$ = 10

$\sigma = \sqrt{10} = 3.16$

$Z = \frac{15.5 - 10}{3.16\ 6} = \frac{5.5}{3.16} = 1.71$

The probability for Z = 1.74 = 0.4591

p(Z≥ 1.74) = 0.5000 – 0.4591 = 0.0409

### Check Your Progress –8

Patients arrive at a hospital at an average rate of 25 per a day. What is the probability that more than 22 patients will arrive in a day? Assuming arrival of patients follow the poisson distribution

## 3.5 ANSWERS TO CHECK YOUR PROGRESS

1. P(E) = 0.5

   P(B) = 0.5

| X | P(x) |
|---|---|
| 0 | 0.0625 |
| 1 | 0.250 |
| 2 | 0.375 |
| 3 | 0.250 |
| 4 | 0.0625 |
| | Sum 1 |

2. $\sigma^2 = 0.27$

   $\sigma = 0.5196$

3. p = 0.60 q = 0.4 n = 6

   a)

| X | p(x) |
|---|---|
| 0 | 0.0041 |
| 1 | 0.03684 |
| 2 | 0.13824 |
| 3 | 0.27648 |
| 4 | 0.31104 |
| 5 | 0.1866 |
| 6 | 0.04666 |
| | 1 |

   b) E(x) = np = 3.6

   Standard deviation = $\sqrt{npq}$

   = $\sqrt{3.6(0.4)} = \sqrt{1.44}$ = 1.2

4. 1) N = 50

   S = 40

   n = 5

   x = 4 P(x = 4) = 0.4313

5. a) The company will meet its goal if line failures do not exceed three so the probability that the company will meet its goal is

   P 6 + 1 + 2 + 3 line failures

   = 0.4335

   b) p(will not meet its goal)

   = 1 – p (will meet its goal)

   = 1 – 0.4335

   = 0.5665
6. a) 0.7029

   b) 0.3821

   c) 38350 lm
7. $\sigma$ = 0.005 mm
8. P(x > 22) = 0.6915

## 3.6 MODEL EXAMINATION QUESTIONS

Answer the following questions (clearly show your steps)

1. List the characteristics of the normal or continuous probability distribution and its accompanying normal curve.
2. Why we apply the normal probability distribution.
3. What determines the shape of the normal curve why?
4. Service life of truck tyres for heavy-duty trucks follows the normal distribution with mean 50,000km and standard deviation 5000km.

   a) Calculate Z value for 60,000km, 48,000km, 63,000km, 58,000km, 39,000km, 62,750km.

   b) What is the probability that a tyre will last

   i) between 47,000km and 50,000km?

   ii) between 50,000 and 60,000km

   iii) between 45,000 and 57,500km?

   iv) less than 48,000km?

   v) greater than 45,000km?

vi) less than 63000km?

vii) between 53,000 and 62,000km?

viii) between 55,000 and 63,000km?

c) The supplier of the tyres is planning to replace only 1% of those tyres with the least performance. What should be the service life for warranty?

d) Tyres with less than 38500km performance are considered below standards or defective. How many tyres will be below standards, if 2500 tyres are made?

5. Sales at a department store follow the normal distribution with an unknown mean and unknown standard deviation. The retailing manager does know, however that, 16% of the time he sells more than 2200 assortments and 34% of the time he sells less than 1800 assortments. Find the mean and standard deviation for the number of items sold.
6. For an Airline 80% of the time seats in all flights are occupied. If a particular Air plane has 180 seals
   a) What is the expected number of occupied seats?
   b) What is the probability (applying normal approximation to the binomial) that
      i. More than 150 seats will be occupied
      ii. Less than 175 seats will be occupied
      iii. 190 or more seats will be occupied
7. Customers arrivals at a bank follow the poisson distribution with an average rate of 45 in an hour. What is the probability that in a particular one hour time
   a) more than 50 will arrive
   b) 55 or more will arrive
   c) 35 to 55 will arrive

## UNIT 4: SAMPLING AND SAMPLING DISTRIBUTION

### Contents


4.0 Aims and Objectives
4.1 Introduction
4.2 Why Sampling
4.3 Errors
4.4 Probability Sampling
4.5 Method of Probability Sampling
4.6 Sampling Distribution
4.7 Central Limits Theorem
4.8 Distribution of the Standardized Statistics
4.9 Estimates
4.9.1 Point Estimates and their Properties
4.9.2 Interval Estimates
4.9.2.1 Constructing Confidence Interval
4.9.2.2 Finite Population Correction Factor
4.10 Selecting A Sample Size
4.10.1 Sample Size for the Mean
4.10.2 Sample Size for Proportion
4.11 Answers to Check Your Progress
4.12 Model Examination Question


## 4.0 AIMS AND OBJECTIVES

Usually the population under study is very large or infinite which makes studding it very difficult or impossible. Under such circumstances we take a sample or a subset of the population to study the population. After completing this unit, you will be able to

- ☞ understand why we sample
- ☞ identify types of probability sampling techniques
- ☞ define sampling distribution and the central limit theorem

- ☞ estimate the population mean and population proportion
- ☞ identify the types of estimates and construct confidence interval for the mean and proportion
- ☞ determine the sample size for the mean and the proportion

## 4.1 INTRODUCTION

Statistics is a science of inference. It is the science of making general conclusion about the entire group (the population) based on information obtained from a small group or sample.

## 4.2 WHY SAMPLING

It is often not feasible to study the entire population. The following are some of the major reasons why sampling is necessary.

### 4.2.1 The Destructive Nature of Certain Testes

Many experiments especially in quality control demand destructing outputs. Consider the following tests:

- Testing wine or coffee
- Blood test for a patient
- Testing strength of light bulbs
- Seed test for germination etc.

Unless sample is taken from the entire population the wine tester should drink all the wine, all the blood from the patient should be poured-out, all the light bulbs produced should be destroyed and nothing would remain for sale. Here sample is a must.

### 4.2.2 The Physical Impossibility of Checking all Items in the Population

The populations of fish, birds and other wild lives are large and are constantly moving being born and dying. There is no mechanism to contact all items or individual members of the population.

### 4.2.3 The Cost of Studying all the Items in a Population is Often Prohibitive

Public opinion polls and consumer testing organizations usually contact fewer families out of millions. Consider a multi national corporation with 50 million customers world wide. If this company plans to undertake market survey out of the 50 million it will take 2000 samples, if it takes 20 br. to mail samples and tabulate the responses of 2000 samples, total survey will cost Br. 40000. While the same survey involving 50 million population would cost about one billion br.

### 4.2.4 The Adequacy of Sample Results

Even if funds were available, it is doubtful whether the additional accuracy of 100% sample i.e., studying the entire population is essential in most problems. To determine monthly index of food prices, bread, beans, milk etc, it is unlikly that the inclusion of all grocery stores and shops would significantly affect the index, Since, the prices of such commodities usually do not vary by more than a few cents form one store to another. 100% accuracy cannot be all ways guaranteed by studying the entire population. The chance of error in collecting and analyzing bulk data has its own disadvantage.

### 4.2.5 To Contact the Whole Population Would Often be Time Consuming

A market survey may take two or three days for field interviews by taking a sample of 2000 customers. By using the same staff and interviewers and working seven days a week it would take nearly 200 years to contact 50 million customers.

## 4.3 ERRORS

Avery important consideration in sampling is to select the sample in such a way that it is very likely to have characteristics similar to the population as a whole. Other wise, the sample could have characteristics quite different form the population. In that case you could draw erroneous conclusions about the population on the basis of improperly chosen sample. Error can be sampling or non-sampling error.

Sampling error is related with the sampling technique and approaches while non-sampling error is related with administering the survey. Sampling errors can be identified and rectified using some mathematical techniques. While the non-sampling errors are very difficult to identify and rectify before making conclusions.

## 4.4 PROBABILITY SAMPLE

Probability sample is a sample selected in such away that each item or person in the population being studied has a known (nonzero) likelihood of being included in the sample. Non-probability sample is a sample selected based on contingency and judgment.

If non-probability methods are used, not all items or people have a chance of being included in the sample. In such instances the result may be biased, the sample result may not be representative of the population.

Panel sampling and convenience sampling are non-probability sampling. They are based on convenience to the statistician. Statistical procedures used to evaluate sample results based on probability sampling.

## 4.5 METHODS OF PROBABILITY SAMPLING

All probability sampling methods have one goal, to allow chance to determine the items or persons to be included in the sample. There are different types of sampling techniques. However there is no one best method of selecting a probability sample. A technique best for a given circumstance or situation may fail in another situations.

Commonly used probability sampling techniques are the following:

### 4.5.1 Simple Random Sampling

A sample formulated in such a manner that each item or person in the population has the same chance of being included in the sample. We can easily list the name or identification of all items i.e. the population on a piece of paper and properly fold and mixing and ruing the lot until we have the required sample size. This method is time consuming and awkward.

More convenient method of selecting a random sample is to use a table of random numbers. It is necessary first to give identification for all elements in the population. We will select the starting point arribitrarily and continue to take the sample until we have the required sample size.

This method may be to use in certain research situations. Mostly difficult when the population is very larger.

### 4.5.2 Systematic Random Sampling

The items or individuals of the population are arranged in some way (alphabetical) or some other method. A random starting point is selected and then every $K^{th}$ member of the population is selected for the sample.

A systematic random sample should not be used, if there is a predetermined pattern to the population. Like inventory control, or if values are listed in ascending or descending orders.

### 4.5.3 Stratified Random Sample

A population is first divided into subgroups called strata, and a sample is selected form each stratum. Stratum can be

- Proportional sample / to the population or
- Non-proportional

**Example.** Studying advertising expenditure of 352 large companies. Profitability percentage is used to stratify this population. We need to select 50 samples.

| **Stratum (0)** | **Profitability (1)** | **Number of (2)** | **% of total (3)** | **Number (4) (50x(3))** |
|---|---|---|---|---|
| 1 | 30 % and over | 8 | 2 | 1 |
| 2 | 20-30% | 35 | 10 | 5 |
| 3 | 10-20% | 189 | 54 | 27 |
| 4 | 0 up to 10% | 115 | 33 | 16 |
| 5 | deficit | 5 | 1 | 1 |
| | | **352** | **150** | **50** |

Stratified sampling has the advantage, in some cases, of more accuracy reflecting the characteristics of the population than dose simple random or systematic random sampling.

### 4.5.4 Cluster Sampling

It is dividing the population in to small units. These units are called primary units. There select at random certain groups or clusters. This technique is

Often employed to reduce cost of sampling a population scattered over a large geographic area.

## 4.6 SAMPLING DISTRIBUTION

Two important terms in sampling distributive:

a) Population parameter – A numerical measure of a population, population mean, $\mu$ population variance, $\sigma^2$, population standard deviation, $\sigma$, population proportion, p etc.
b) Sample statistics / Statistic/ - A numerical measure of the sample

Sample mean, $\bar{x}$, sample variance $S^2$ sample standard deviation S, sample proportion $\bar{p}$, etc.

### Sampling Distribution of the means ($\bar{x}$)

Sampling distribution of the sample means, $\bar{x}$, is the probability distribution consisting of a list of all possible sample means of a given sample size selected from a population, and the probability of occurrence associated with each sample mean.

**Example.** The following distribution is the hourly wage of seven employees

| Employee | Hourly wage |
|---|---|
| A | 7 |
| B | 7 |
| C | 8 |
| D | 8 |
| E | 7 |
| F | 8 |
| G | 9 |

This population has a mean of 7.71 hoary wage i.e. $^{54}/_{7}$

If we are planning to take sample of two employees, we will have 21 ($_7C_2$) possible samples and corresponding sample means. The 21 possible samples with their mean are the following:-

| **Possible Sample** | **Sample mean ($\bar{x}$)** |
|---|---|
| AB | 7.0 |
| AC | 7.5 |
| AD | 7.5 |
| AE | 7.0 |
| AF | 7.5 |
| AG | 8.0 |
| BC | 7.5 |
| BD | 7.5 |
| BE | 7.0 |
| BF | 7.5 |
| BG | 8.0 |
| CD | 8.0 |
| CE | 7.5 |
| CF | 8.0 |
| CG | 8.5 |
| DE | 7.5 |
| DF | 8.0 |
| DG | 8.5 |
| EF | 7.5 |
| EG | 8.0 |
| FG | 8.5 |
| | $\Sigma\bar{x}$= 162 |

Summary of sampling distribution of the means for n=2 will be

| Sample mean | No of means | Probability |
|---|---|---|
| 7 | 3 | 0.1429 |
| 7.5 | 9 | 0.4285 |
| 8.00 | 6 | 0.2857 |
| 8.50 | 3 | 0.1429 |
| Total | 21 | 1 |

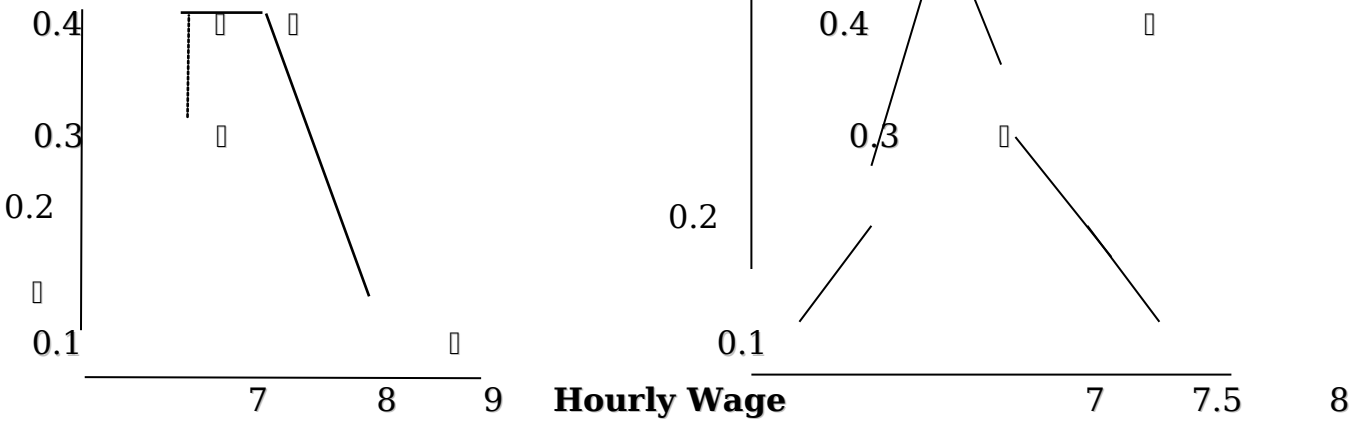
The mean of the distribution of sample means is obtained by summing the various sample means and dividing the sum by the number of samples. The mean of all the sample means is usually written $\mu_{\bar{x}}$ $\mu$ reminds us that it is a population value because we have considered all possible samples. The subscript $\bar{x}$ indicates that it is a sampling distribution of means.

$$\mu_{\bar{x}} = \frac{7+7.5+\ldots+8.5}{21} = \frac{162}{21} = 7.71$$

The following graphs represent the population distribution and the distribution of the sample means.

**Population Distribution** **Probability**

**Sampling Distribution**



From the above graphs / distributions we can understand that:

a) The mean of the sample means (7.71) is equal to the mean of the population. This is always true if all possible samples of a given size are selected from the population of interest
b) The range of sample means is less than the range in the population. The sample means range form 7 to 8.5 where as the population vary form 7 to 9.00.
c) The graph representing the distribution of the population and that of the sample means shows the change in shape from the population to the sample. The graph representing the distribution of the sample means looks like a normal curve.

## 4.7 THE CENTRAL LIMIT THEOREM

For a population with mean $\mu$ and Variance $\sigma^2$, the sampling distribution of the means of all possible samples of size n generated from the population will be approximately normally distributed with the mean of the sampling distribution equal to $\mu$ and the variance equal to $\sigma^2/n$, assuming that the sample size is sufficiently large.

The important facets of the central limit theorem bear repeating.

1. if the sample size n is sufficiently large, the sampling distribution of the means will be approximately normal regardless of the distribution of the population form which the random sample is drawn
2. if a population is large and a large number of samples are selected from the population then the means of the sample means will be close to the population mean.
3. the variance of the distribution of sample means is determined by $\sigma^2/n$. This implies that as the sample size increases the variation of $\bar{x}$ about its mean decrease. Note that a sample of

30 or more elements is considered sufficiently large for the central limit theorem to take effect.

A larger minimum sample size may be required for a good normal approximation when the population distribution is very different from a normal distribution. While a smaller minimum sample size may suffice for a good normal approximation when the population distribution is close to a normal distribution.

## 4.8 DISTRIBUTION OF THE STANDARDIZED STATISTICS FOR THE SAMPLE MEAN

In order to use the central limit theorem, we need to know the population standard deviation when it is not know the standard deviation of the sample, designated by S is used to approximate it. The standardized distribution of the sample means is Z and

$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$, if the population standard deviation is known or $\frac{\bar{x} - \mu}{s/\sqrt{n}}$, if the population standard deviation is unknown.

**Example 1:**

The annual wages of all employees of a company has a mean of 20,400 per year with standard deviation of 3200. The personnel manager is going to take a random sample of 36 employees and calculate the sample mean wage. What is the probability that the sample mian will exceed 21.000?

n= 36 $\mu$ = 20,400 and $\sigma$ =3200

$P[\bar{x} > 21{,}000] = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{21000 - 20400}{3200/\sqrt{36}} = 1.125$

$P(Z > 1.13) = 0.1292$

**Example. 2**

A company makes engine used in speedboats. The company's engineers believe that the engine delivers an average power of 220 horse power / HP/ and that the standard deviation of power delivered is 15 HP. A potential buyer intends to sample 100 engines (each engine to be run a single time ) . What is the probability that the sample mean, $\bar{x}$, will be less than 217 HP.

$$P(\bar{x} < 2/7) = P\left(Z < \frac{217 - \mu}{\sigma/\sqrt{n}}\right) = \frac{217 - 220}{15/\sqrt{100}} = -2$$

$$P(Z < -2) = 0.0228$$

Thus if the population mean is indeed $\mu$ = 220 HP and the standard deviation is $\sigma$ = 15 HP, there is a rather small probability that the potential buyer's tests will result in a sample mean lower than 217HP.

## Check Your Progress –1

The average GPA of all graduating students in a college is 2.85 with a standard deviation of 0.96. The placement unit randomly selects 64 graduating students. What is the probability that the sample mean will be greater than 3.00?

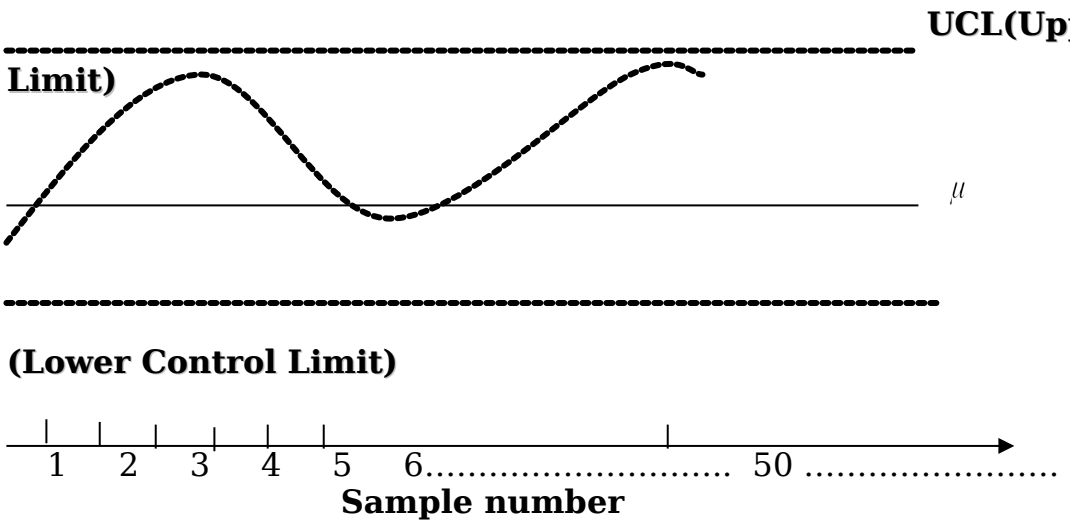
One important application of the central limit theorem is in the area of quality control. The manufacturing process is variable and be monitored to be sure that the variability does not get beyond acceptable levels.

A control chart is used to assist in monitoring the variability $\bar{x}$ chart is used to control variation in the sample means.

The Chart has two limits about the mean $\mu$

c) Upper control limit (UCL)
d) Lower control limit (LCL)

The centerline is the desired mean, $\mu$.




If a point is observed above UCL or below LCL the process is stopped and find the problem.

The upper and lower control limits are generally located one, two, or three times $\sigma_{\bar{x}}$ above and below μ depending on the nature of the product and the process.

## 4.9 ESTIMATES

Inferential statistics is concerned with estimation.

In many cases values for a population parameter are unknown. If parameters are unknown it is generally not sufficient to make some convenient assumption about their values, rather those unknown parameters should be estimated.

In business many decision are made with out complete information.

A firm does not know exactly what will be its sales volume next year or next month. A college does not know exactly how many students will enroll next year. Both must estimate to make decision about the future.

***Types of Estimates***

## 4.9.1 Point estimate

A number or a simple number is used to estimate a population parameter. A random sample of observations is taken from the population of interest and the observed values are used to obtain a point estimate of the relevant parameter.

a. The ample mean, $\bar{x}$, is the best estimator of the population mean $\mu$. Different samples from a population yield different point estimates of $\mu$,

b. Sample proportion $\bar{p}$ is a good estimator of population proportion, p.

- Population proportion P is equal to the number of elements in the population belonging to the category of interest divided by the total number of elements in the population p = $\frac{X}{N}$

Where: X is the number of success in the population and

N population size

Sample proportion, $\bar{p} = \frac{x}{n}$ where;

x is the number of elements in the sample found to belong to the category of interest and n is the sample size.

or $\bar{p}$ = Number of success in a sample / number sampled

**Example** of 2000 persons sampled 1600 favored more strict environmental protection measures, what is the estimated population proportion.

$\bar{p} = \frac{16000}{2000} = 0.80$

80% is an estimate of the proportion in the population that favor more strict measures

In general:

The statistic $\bar{x}$ estimates $\mu$

S estimates $\sigma$

$S^2$ estimates $\sigma^2$

$\bar{p}$ estimates p

***Estimators and their properties / Goodness of an estimator/***

The properties of good estimators are

a) Un biasedness
b) Efficiency
c) Consistency and
d) Sufficiency

a) An estimator is said to be unbiased if its expected value is equal to the population parameter it estimates.

$E(\bar{x}) = \mu$ The sample mean , $\bar{x}$, is therefore, an unbiased estimator of the population mean. Any systematic deviation of the estimator away from the parameter of interest is called Bias.

b) An estimator is efficient if it has a relatively small variance (as standard deviation). The sample means have a variance of $\sigma/n$ value is less than $\sigma$. So the sample mean is an efficient estimator of the population mean.

c) An estimator is said to be consistent if its probability of being close to the parameter it estimates increases as the sample size increases.

The sample mean is a consistent estimator of $\mu$. This is so because the standard deviation of $\bar{x}$ is $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. As the sample size n increases, the standard deviation of $\bar{x}$ decreases and hence the probability that $\bar{x}$ will be closes to its expected value, $\mu$, increases.

d) An estimator is said to be sufficient if it contains all the information in the data about the parameter it estimates. The sample mean is sufficient estimator of μ. Other estimators like the median and mode do not consider all values. But the mean considers all values (added and divided by the sample size).

### 4.9.2 Interval Estimates

Interval estimate states the range within which a population parameter probably lies. The interval with in which a population parameter is expected to lie is usually referred to as the *confidence interval*.

The confidence interval for the population mean is the interval that has a high probability of containing the population mean, μ

Two confidence intervals are used extensively.

1. 95% confidence interval and
2. 99% confidence interval

A 95% confidence interval means that about 95% of the similarly constructed intervals will contain the parameter being estimated. If we use the 99% confidence interval we expect about 99% of the intervals to contain the parameter being estimated.

Another interpretation of the 95 % confidence interval is that 95 % of the sample means for a specified sample size will lie with in 1.96standred deviations of the hypothesized population mean. For 99% the sample means will lie, with in 2.58 standard deviations of the hypothesized population mean.

Where do the values 1.96 and 2.58 come form?

The middle 95% of the sample mean lie equally on either side of the mean. And logically 0.95/2=0.4750 or 47.5% of the area is to the right of the mean and the area to the left of the mean is 0.4750.

The Z value for this probability is 1.96.

The Z to the right of the mean is + 1.96 and Z to the left is – 1.96.

### 4.9.2.1 Constructing Confidence Interval

a) Compute the standard error of the mean

Standard error of the mean is the standard deviation of the sample means.

$\sigma$ = population standard deviation

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

If the population standard deviation is not know, the standard deviation of the sample s, is used to approximate the population standard deviation.

$$S_{\bar{x}} = \frac{S}{\sqrt{n}}$$

This indicates that the error in estimating the population mean decreases as the sample size increases.

b) The 95% and 99% confidence intervals are constructed as follows when n $\geq$ 30.

95% confidence interval $\bar{x} \pm 1.96 \frac{S}{\sqrt{n}}$

99% confidence interval $\bar{x} \pm 2.58 \frac{S}{\sqrt{n}}$

1.96 and 2.58 indicate the Z values corresponding to the middle 95% or 99% of the observation respectively.

In general a confidence interval for the mean is computed by $\bar{x} \pm Z\frac{S}{\sqrt{n}}$, Z reflects the selected level of confidence.

**Example.** An experiment involves selecting a random sample of 256 middle managers for studying their annual income. The sample mean is computed to be Br. 35,420 and the sample standard deviation is Br. 2,050.

a. What is the estimated mean income of all middle managers ( the population ) ?
b. What is the 95% confidence interval c(rounded to the nearest 10)
c. What are the 95% confidence limits?
d. Interpret the finding.

**Solution**

a. Sample mean is 35 420 so this will approximate the population mean so $\mu$ = 35420. It is estimated from the sample mean.
b. The confidence interval is between 35170 and 35670 found by

$$\bar{X} \pm 1.96\frac{S}{\sqrt{n}} = 35420 \pm 1.96\left(\frac{2050}{\sqrt{256}}\right) = 35168.87 \text{ and } 35671.13$$

c. The end points of the confidence interval are called the confidence limits. In this case they are rounded to 35170 and 35670. 35170 is the lower limit and 35070 is the upper limit.
d. Interpretation

If we select 100 samples of size 256 form the population of all middle managers and compute the sample means and confidence intervals, the population mean annual income would be found in about 95 out of the 100 confidence intervals. About 5 out of the 100 confidence intervals would not contain the population mean annual income.

## Check Your Progress –2

A research firm conducted a survey to determine the mean amount smokers spend on cigarette during a week. A sample of 49 smokers revealed that the sample mean is Br. 20 with standard deviation of Br. 5. Construct 95% confidence interval for the mean amount spent.

### *Confidence interval for a population proportion*

The confidence interval for a population proportion is estimated

$\bar{p} \pm Z\sigma_p$

Where $\sigma_p$ is the standard error of the proportion and

$$\sigma_{\bar{p}} = \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

Therefore the confidence interval for population proportion is constructed by

$$\bar{p} \pm Z \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

**Example.** Suppose 1600 of 2000 union members sampled said they plan to vote for the proposal to merge with a national union. Union by laws state that at least 75% of all members must approve for the merger to be enacted. Using the 0.95 degree of confidence, what is the interval estimate for the population proportion? Based on the confidence interval, what conclusion can be drawn? $\bar{p} = \frac{1600}{2000} = 0.8$. The sample proportion is 80%

The interval is computed as follows. $\bar{p} \pm Z \sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = 0.80 \pm 1.96 \sqrt{\frac{0.80(1-0.8)}{2000}} = 0.08 \pm 1.96 \sqrt{0.00008}$

= 0.78247 and 0 – 81753 rounded to 0.782 and 0.818.

Based on the sample results when all union members vote, the proposal will probably pass because 0.75 lie below the interval between 0.782 and 0.818.

## Check Your Progress –3

A sample of 200 people were assumed to identify their major source of news information; 110 stated that their major source was television news coverage. Construct a 90% confidence interval for the proportion of people in the population who consider television their major source of news information.

### 4.9.2.2 Finite Population Correction Factor

The population we have sampled so far has been very large, or assumed to be infinite.

If the sampled population is not infinite or not larger we need to make some adjustments in the standard error of the mean and the standard error of the proportion. This is done to reduce the error we committee in estimating a parameter.

A population that has a fixed upper bond is said to be finite. A finite population can be small or can be very large.

For a finite population, where the total number of objects is N, and the size of the sample is n the following adjustment is made to the standard errors of the mean and the proportion.

Standard error of the mean

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

**Standard error of the proportion**

$$\sigma_{\bar{p}} = \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}\sqrt{\frac{N-n}{N-1}}$$

This adjustment is called finite population correction factor.

Why is it necessary to apply a factor and what is its effect?

Logically, if a sample is a substantial percentage of the population, then we would expect any estimate to be more precise than those for a smaller sample.

Suppose the population is 1000 and the sample is 100. Then this ratio is $\frac{1000-100}{1000-1}$ or $\frac{900}{999}$. Taking the square root gives the correction factor 0.9492. Multiplying the standard error reduces the error by about 5% or (1-0.9492)= 0.5. This reduction of the size of the standard error yields a

smaller range of values in estimating the population mean. If the sample size is 200 the correction factor is 0.8949. Meaning that the standard error has been reduced by more than 10%.

The usual rule is that If the ratio of the sample to the population, $^{n}/_{N,}$ is less than 0.05, the finite population correction factor is ignored.

**Example**. There are 250 families in a small town A poll of 40 families revealed that the mean annual community contribution is 450 with a standard deviation of 75. Construct a 95% confidence interval for the mean annual contribution.

***Solution: -***

First note that the population is finite.

Second the sample constitute more than 5% of the population n/N = 40/250 =0.16 Hence the finite population correction factor is applied.

$$\bar{x} \pm Z\frac{S}{\sqrt{n}}\left(\sqrt{\frac{N-n}{N-1}}\right) = 450 \pm 1.96\left(\frac{75}{\sqrt{40}}\right)\sqrt{\frac{250-40}{250-1}}$$

$$= 450 \pm 23.24\ \sqrt{0.843}$$

$$= 450 \pm 21.34$$

$$= 428.66 \text{ and } 471.34$$

## Confidence interval for small sample (Student's Distribution)

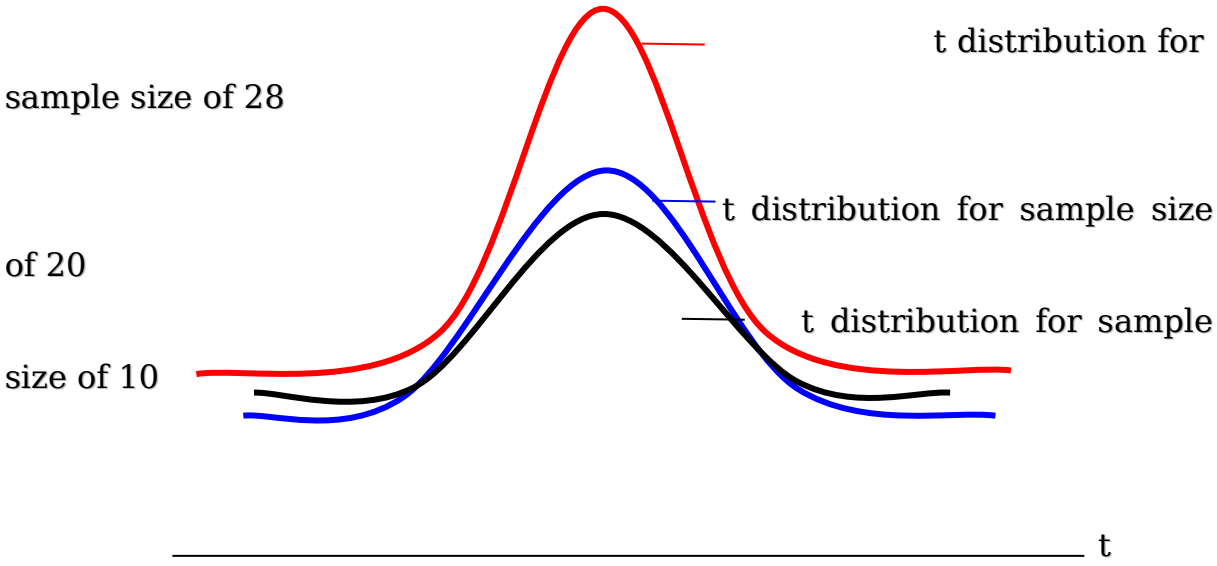
When the population is large and normal and the standard deviation is known the standard normal distribution is employed to construct the confidence interval for the mean and proportion. If the sample size is at least 30, the sample standard deviation can substitute the population standard deviation and the results are deemed satisfactory.

If the sample size is less than 30 and population standard deviation is unknown, the standard normal distribution, Z, is not appropriate. The student's t or the t distribution is used.

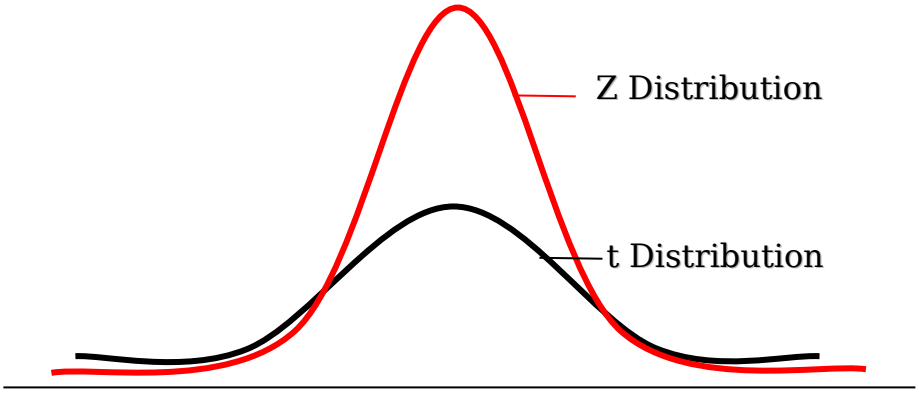
### Characteristics of the Student's t Distribution

Assuming that the population of interest is normal or approximately normal, the following are the characteristics of the t distribution

1. It is a continuous distribution
2. It is bell-shaped and symmetrical
3. There is not one t distribution, but rather a 'family' of t distribution. All have the same mean of zero but their standard deviation differ according to the sample size, n. The t distribution differs for different sample size.
4. It is more spread out and flatter at the center than is the Z. However as the sample size increases the curve representing t distribution approaches the Z distribution.



As the sample size decreases the curve representing the t distribution will have wider tails and will be more flat at the center.




For a given confidence level, say 95%, the t value is greater than the Z value. This is so because there is more variability in sample means computed from smaller samples. Thus our confidence in the resulting estimate is not strong. t values are found referring to the appropriate degrees of freedom in the t table. Degrees of freedom means the freedom to freely move data points or the freedom to freely assign values arbitrarily.

Degrees of freedom (df) = n – 1 where n is the sample size.

This implies that we can freely move or assign values for all data points except the last $n^{th}$ value. If the mean of the distribution is specified there is a freedom to assign any value for all data points except the lost point.

**Example -** the mean of five data points is 12. Then it follows that the sum of all the five points is 60 = (5 x 12). Thus if five points are constrained to have a sum of 60 or a mean of 12, we have 5 – 1 = 4 degrees of freedom.

If all the five data points are missing we are free to assign any value as long as their sum is 60 say 14, 12, 10, 9, 15.

If 4 are missing we are free to assign any value since 60 minus the known value of a data point is known.

If two are un know, 14, 16, 10, $x_3$, $x_4$ since $14 + 16 + 10 + x_3 + x_4 = 60$

Then $x_3 + x_4 = 60 - 40 = 20$

$x_3 + x_4 = 20$. We can assign any value as long as their sum is 20. 10, 10 or 9.11 or 15.5 etc…

But if the four data points are known, (10, 14, 16, 12), the $5^{th}$ data point will have a predetermined value i.e. 60 – 52 = 8. Now we are not free to assign arbitrary value for this data point. Degrees of freedom can be obtained from the deviation based on the assumption that sum of the differences (d) between the mean and all values of the random variable (x) is zero. I.e., if we subtract the mean from all values of x the sum of the difference will be zero consider the above five data points. Their mean is 12 and their sum 60. Thus $(x_1 - 12) + (x_2 + 12) + (x_3 - 12) + (x_4 - 12) + (x_5 - 12) = 0 = d_1 + d_2 + d_3 + d_4 + d_5 = 0$

Now we are free to assign any value for only four missing differences as long as this sum is zero. So we have still n – 1 degrees of freedom.

**Computing t value**

The t variable representing the student's t distribution is defined as

$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$ where: $\bar{x}$ is the sample mean of n measurements, $\mu$ is the population mean and s is the sample standard deviation

Note that t is just like $Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ except that we replace $\sigma$ with s. unlike our methods of large samples, $\sigma$ cannot be approximated by s when the sample size is less than 30 and we can not use the normal distribution. The table for the t distribution is constructed for selected levels of confidence for degree of freedom up to 30. To use the table we need to know two numbers, the tail area, (1 minus confidence level selected), and the degree of freedom.

(1 – confidence level selected) is $\alpha$, the Greek letter alpha. This is the error we committee in estimating.

The confidence interval for the sample mean is $\bar{x} \pm {}^{\alpha}/_{2}(n-1)\,{}^{S}/_{\sqrt{n}}$

**Example**. A traffic department in town is planning to determine mean number of accidents at a high-risk intersection. Only a random sample of 10 days measurements were obtained.

Number of accidents per day were

8, 7 10 15 11 6 8 5 13 12

Construct a 95% confidence interval for the mean number of accident per day.

a) Compute $\bar{x}$ and s

$\bar{x} = \frac{95}{10}$ = 9.5 per day

$S_x = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}} = \sqrt{\frac{945}{9}}$ = 3.24 per day

The confidence level is 95% so

$\alpha = 1 - 0.95 = 0.05$

$\frac{\alpha}{2} = \frac{0.05}{2}$ = 0.025

The degree of freedom, df = n – 1 = 10 – 1 = 9 from the t table $t_{0.025}$, df 9 = 2.76

The confidence interval is

$\bar{x} \pm t_{.0025}$ df(9) $\frac{s}{\sqrt{n}}$

$9.5 \pm (2.26)\ \frac{3.24}{\sqrt{10}}$

9.5 ± 2.3

7.2 to 11.80

With 95% confidence the mean number of accident at this particular intersection is between 7.2 and 11.8.

## Check Your Progress –4

A quality controller of a company plans to inspect the average diameter of small bolts made. A random sample of 6 bolts was selected. The sample is computed to be 2.0016mm and the sample standard deviation 0.0012mm. Construct the 99% confidence interval for all bolts made.

## 4.10 SELECTING A SAMPLE SIZE

Size of a sample must be determined scientifically. Care must be taken not to select a sample too large or too small. There are two misconceptions about how many to sample

a) Sample consisting 5% (or similar constant percentage) is adequate for all problems.

   5% can be too much for a particular population say 10 million or can be too small for another say 200.

b) A sample, for example, must be selected form a heavily populated area.

The avoid such problems the sample size should be mathematically determined.

### 4.10.1 Sample Size for the Mean

There are three factors that determine the size of the sample. None of which has any direct relation ship to the size of the population.

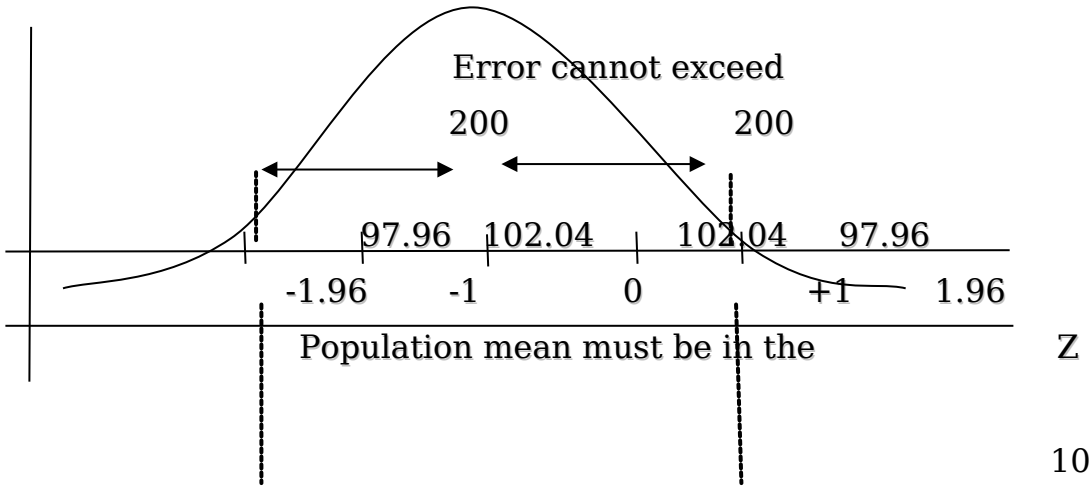
a. The degree of confidence selected.
b. The maximum allowable error
c. The variation in the population

a. *The degree of confidence,* This is usually 95% or 99%. But it may be any level. It is specified by the statistician. The higher the degree of confidence, the larger the sample required. If we want to be sure the true mean will lie between an interval, we would hve to survey the entire population. Example. Suppose the parameter to be estimated is the arithmetic mean, and the degree of confidence selected is 90%. Based on a sample, it was estimated that the population mean is in the

interval between 850 and 1050. Logically, if the degree of confidence were increased to 95% or 99% the sample size would have to increase.

b. *Maximum error allowed*. It is the maximum error that will be tolerable at a specified level of confidence. Suppose a statistician is interested to estimate the mean income of residents of an area. There are indications that the family incomes range from a probable low of 19000 to a high of about 39000. On the assumption that these are reasonable estimates , does it seem likely that the statistician would be satisfied with this statement resulting from a sample of area residents. " The population mean is between 23,000 and 35,000" Probability not. Because confidence limits that wide indicate little or nothing about the population mean. Instead, the statistician stated "using the 0.95 confidence level, the total error is predicting the population mean should not exceed by 200". The maximum allowable error is denoted 'E' = E = $|\bar{x} - \mu|$. This means based on a sample size n, if the estimate of population mean is computed to be 35,000, then we will assure that the population mean is in the interval between 34800and 35200. Found by 35,000 + 200 and 35000-200. For the 0.95 degree of confidence selected the maximum error of $\pm$ 200 interms of Z is 1.96. To determine the value of one standard error of the mean $\sigma_{\bar{x}}$ simply divide the total error of 200 by 1.96 = 102.04

$$\sigma_{\bar{x}} = \frac{200}{1.96} = 102.04$$



interval ± 200 from the sample mean

The size of the sample is computed by solving for n in the formula $S_{\bar{x}} = \frac{S}{\sqrt{n}}$, note that since we are using a sample standard deviation.

i.e., $S_{\bar{x}}$ is substituted for $\sigma_{\bar{x}}$ and S for σ

$$\frac{Total\ allowable\ error}{Z\ standard\ deviation} = \frac{Sample\ standard\ deviation}{\sqrt{sample\ size}}$$

Total allowable error let be represented by 'E' then it follows that,

$$\frac{E}{Z} = \frac{S}{\sqrt{n}} = \frac{200}{1.96} = \frac{S}{\sqrt{n}} = 102.04$$

Since there are two unknowns for one equation we cannot solve for both.

c. *Variation in the population*. There are still two unknowns. To solve for the number to be sampled we need to estimate the variation in the population. The standard deviation is a measure of variation. Thus the standard deviation of the population must be estimated.

This can be done either:

a- By taking a small pilot survey and using the standard deviation of the pilot sample as an estimate of the population standard deviation or

b- By estimating the standard deviation based on knowledge of the population.

Suppose a pilot survey is conducted and sample standard deviation is computed to be 3000. The number to be sampled can now be estimated.

$$S_{\bar{x}} = \frac{S}{\sqrt{n}} = \frac{E}{Z}$$

$$\frac{200}{1.96} = \frac{3000}{\sqrt{n}} \qquad n = 864.36$$

$S_{\bar{x}}$ is standard error of the mean, the error we commit in estimating . From the above computation we can learn that as the variation in the population increase the sample size will increase.

A more convenient computational formula for determining n is.

$$n = \left(\frac{Z.S}{E}\right)^2$$

where E = allowable error

Z = Z value for the degree of confidence selected

S = Sample deviation

*For this example* $n = \left(\frac{1.96 \times 3000}{200}\right)^2 = 864.36$

**Example 1**. A marketing research firm wants to conduct a survey to estimate the average amount spent on entertainment by each person visiting a popular pub. The people who plan the survey would like to be able to determine the average amount spent by all people visiting the pub to within br. 120, with 95% confidence. From past operations of the pub, an estimate of the population standard deviation is σ = br. 400 what is the minimum required sample sizes?

Z = 1.96

E = 120

σ = 400

Required, n?

$$n = \left(\frac{1.96 \times 400}{120}\right)^2 = 42.68 \approx 43$$

## Check Your Progress –5

A processor of carrots cuts the green top of each carrot washes the carrots, and inserts six to a package. Twenty packages are inserted in a box for shipment. To test the Wight of the boxes, a few were checked. The mean weight was 10kg and the standard deviation 0.25kg. How many boxes must the processor sample to be 95% confident that the sample mean does not differ from the population mean by more than 0.1 kg?

### 4.10.2 Sample size for proportion

The procedure used to determine the sample size for the mean is applicable to determine when proportions are involved.

Three things must be specified.

- Decide on the level of confidence
- Indicate how precise the estimate of the population proportion must be
  - Approximate the population proportion, P, either from past experience or from a small pilot survey $\bar{p}$

The formula for determining the sample size n for a proportion

$$n = \bar{p}(1 - \bar{p}) \left(\frac{Z}{E}\right)^2$$

where: $\bar{p}$- estimated proportion

Z = Z value for the selected confidence level

E = the maximum tolerable error

**Example 1.** A member of parliament wants to determine her popularity in her a region. She indicates that the proportion of voters who will vote for her must be estimated with in $\pm$ 2 percent of the population proportion. Further, the 95% degree of confidence is to be used. In past elections she received 40% of the popular vote in that area. She doubts whether it has changed much. How many registered voters should be sampled?

Z = 1.96

$\bar{p}$ = 0.40

E = 0.02

$$n = \bar{p}(1 - \bar{p}) \left(\frac{Z}{E}\right)^2$$

$$= 0.40\ (1 - 0.4) \left(\frac{1.96}{0.02}\right)^2 = 2{,}304.96 \approx 2305$$

This sample size might be too large, or
too small or exactly correct depending on the accuracy of $\bar{p}$.

Note: if there is no logical estimate of $\bar{p}$, the sample size can be estimated by letting $\bar{p}$=0.5

**Example 2.** Suppose the president wants an estimate of the proportion of the population that support this current policy on unemployment. The president wants the estimate to be with in 0.04 of the true proportion. Assume a 95% level of confidence and the proportion supporting current policy to be 0.60.

a) How large a sample is required
b) How large would the sample have to be if the estimate were not available?

**Solution:**

a) E = 0.04
Z = 1.96
$\bar{p}$= 0.60

$$n = 0.6(1 - 0.6) \left(\frac{1.96}{0.04}\right)^2$$

$$= 577$$

b) E = 0.4
Z = 1.96
$\bar{p}$= 0.50 (since there is no estimate)

$$n = 0.5\ (1 - 0.5) \left(\frac{1.96}{0.04}\right)^2$$

$$= 600$$

## Check Your Progress –6

The marketing department of a company wishes to study the loyalty pattern of consumers. Loyalty patterns range from extremely loyal to brand snitcher. If the department wishes to estimate the proportion of consumers who are extremely loyal to this brand, what sample size would be necessary to estimate this proportion with 0.05 with 95% confidence?

## 4.11 ANSWERS TO CHECK YOUR PROGRESS

1. 0.1056
2. 18.60 and 21.40
3. 49.21% and 60.79%
4. 1.9996 and 2.0036
5. 24
6. 384

## 4.12 MODEL EXAMINATION QUESTIONS

**Answer the following questions**

1. Explain the central limit theorem and its important facets.
2. An investment consultant reports that the average 12-month return on a random sample of 50 projects was 20.74%. If the standard deviation was 5% for the entire large group of stocks from which the sample of projects was chosen, construct a 95% confidence interval for the average 12-month return for all projects in this group.
3. An advertising executive thinks that the proportion of consumer's who have seen his company's advertisement in newspapers is around 0.65. The executive wants to estimate the customer population proportion to with in ± 0.05 and have a 98% confidence in the estimate. How large a sample should be taken.
4. A company wants to estimate the proportion of its employees, who are satisfied with a new incentive scheme. Out of a total of 1,242

employees, 160 were randomly selected and interviewed. Of the one interviewed, 85 indicated that they were satisfied with the new scheme. Construct a 90% confidence interval for the proportion of all employees who are satisfied with the new decision.

5. A survey is being planned to determine the mean amount of time senior executives watch TV. A pilot survey indicated that the mean time per week is 12 hrs with a standard deviation of 3 hrs. It is desired to estimate the mean viewing time within 0.25 hrs. The 95% degree of confidence is to be used. How many executives should be surveyed?
6. Why sampling?
7. What are the properties of good estimators? Explain
8. A sample of 200 people were asked to identify their major source of news information. 110 said their major source was radio.
    a) Construct a 95% confidence interval for the proportion of people in the population that consider radio their major source of news information
    b) How large a sample would be necessary to estimate the population proportion with a sampling error of 0.05 at 95% confidence.
9. What are the factors that determine the size of the sample?
10. Under what circumstances the finite population correction factor should be applied?
11. The registrar of a college wants to estimate the arithmetic mean final GPA of all graduating senior students. GPAs range between 2.0 and 4.0. The mean GPA is to be estimated with plus and minus 0.05 of the population mean. The 99% confidence is to be used. The standard deviation of a small pilot survey is 0.279.
    How many grade reports (transcripts) should be sampled?
12. In a small town there are 250 families. From 50 families sample 15 regularly attend community meetings. Construct a 95%

confidence interval for the proportion of families attending the meeting regularly.

13. A wine importer needs to report the average percentage of alcohol in bottles of new wine.

    From experience with various kinds of wines, the importer believes the population standard deviation is 1.2%. The importer randomly sampled 60 bottles of the new wine and obtain a sample mean of 9.3%. Give a 90% confidence interval for the average percentage of alcohol in all bottles of the new wine.

14. The manufacturers of a sports car want to estimate the proportion of people in a given income bracket, who are interested in a model. The company wants to know the population proportion to within 0.10 with 99% confidence. Current company records indicate that the proportion may be around 0.25. what is the minimum required sample size for this survey.

15. A survey of a random sample of 1000 managers found that 81% of them had a high need for power. This led to a conclusion that power is a motivator for managers. Construct a 90% confidence interval for the proportion of all managers in the population under study who are motivated by power.

16. The average score of trainees who participated in a special training program is 120 with a standard deviation of 15. A company who sent its employees sampled 36 employees and calculates their mean scores. What is the probability that the sample mean will be less than 115?

17. A business faculty in a university is planning to introduce a new performance evaluation technique. Instructors are required to evaluate their respective department heads. A random sample of 7 instructors from the marketing department was selected and their evaluation recorded. The results were

    72, 81, 69, 78, 80, 75, 79

Construct a 90% confidence interval for the average performance evaluation of all the instructors in the department.

## UNIT 5: TESTS OF HYPOTHESES

### Contents


5.0 Aims and Objectives

5.1 Introduction

5.2 Hypothesis and Hypothesis Testing Defined

5.2.1 Hypothesis

5.2.2 Hypothesis Testing

5.3 Steps for Testing a Hypothesis

5.4 Hypothesis Testing Involving Large samples

5.4.1 Testing for the Population Mean /Large Sample/

5.4.1.1 Population Standard Deviation Known

5.4.1.2 Population Standard Deviation Unknown

5.4.2 Testing for Two Population means

5.4.3 Testing for a Population Proportion

5.4.4 Testing for the Difference between Two Population Proportions

5.5 Hypothesis Testing Involving Small Samples

5.5.1 Characteristics of the student's t Distribution

5.5.2 Test for the Population Mean

5.5.3 Test for Comparison of Two Population Means

5.5.4 Hypothesis Testing Involving Paired Observations

5.6 Testing for Difference of Variance Comparing Two Population Variances

5.7 Answers to Check Your Progress

5.8 Model Examination Questions

## 5.0 AIMS AND OBJECTIVES

When we estimate the value of a parameter we are using methods of estimation. The unknown value of a population parameter is estimated from sample information by constructing confidence interval estimate.

Decision concerning the value of a population parameter are obtained by hypothesis testing, which is the topic of this chapter.
After completing this unit, you will be able to:

- define hypothesis and testing hypothesis
- test hypothesis involving large sample
- test hypothesis involving small sample
- understand the p-value in hypothesis testing
- testing for differences of variance

## 5.1 INTRODUCTION

Most statistical inference centers around the parameters of a population. In hypothesis testing we start with an assumed value of a population parameter. Then a sample evidence is used to decide whether the assumed value is unreasonable and should be rejected, or whether it should be accepted; Hence the statistical inferences made are referred to as hypothesis testing.

## 5.2 HYPOTHESIS AND HYPOTHESIS TESTING DEFINED

### 5.2.1 Hypothesis is a statement or an assumption about the value of a population parameter or parameters.

Examples

- The mean monthly income of all employees of a company is br. 2000.
- The average age of students in a college is 22 years
- 5% of the products of a firm are defective

All these hypothesis have one thing in common:
The population of interest are so large that for various reasons it would not be feasible to study all the items, or persons, in the population

### 5.2.2 Hypothesis Testing Defined

Hypothesis testing is a procedure based on sample evidence and probability distribution used to determine whether the hypothesis is a reasonable statement and should not be rejected, or is unreasonable and should be rejected.

It is simply selecting a sample from the populations, calculate sample statistic and based on certain decision rules accept or reject the hypothesis.

Test statistic is a sample statistic computed from the sample data. The value of the test statistic is used in determining whether or not we may reject the hypothesis.

Decision rule of a statistical hypothesis is rule that specifies the conditions under which the hypothesis may be rejected. We decide whether or not to reject the hypothesis by following the decision rule.

## 5.3 STEPS FOR TESTING HYPOTHESIS

There is a five-step procedure that systematize hypothesis testing.
Hypothesis testing as used by the statisticians does not provide proof that something is true, in the manner in which a mathematician “proves” a statement. It does provide a kind of “proof beyond a reasonable doubt” in the manner of an attorney.

***Step I. Identity the null hypothesis and the alternate hypothesis***

The first step is to state the hypothesis to be tested. It is called the *Null Hypothesis*, designated by $H_o$ and read “H sub-zero”. The capital letter H

stands for hypothesis and the subscript zero implies “no difference or no change. There is usually a ‘not’ or a ‘no’ term in the null hypothesis meaning no change”. The null hypothesis is set up for the purpose of either to rejecting or not to rejecting it. The null hypothesis is a statement that will be rejected it our sample information provide us with convincing evidence that it false. And it will not be rejected if our sample data fail to provide ample evidence that it is false.

If the null hypothesis is not rejected based on sample data, in effect we are saying that the evidence does not allow us to reject it. We cannot state, however, that the null hypothesis is true. This is the same as the situation in the courts.

In courts we heard judges saying, “Found not guilty” when they release a suspect free. They never say “he is innocent”. The suspect is released may be because the prosecutor or the police fail to provide the court with convincing evidence beyond reasonable doubt that the suspect has committed the crime. The null hypothesis is a tentative assumption made about the value of a population parameter. Usually it is a statement that the population parameter has a specific value.

Failure to reject the null hypothesis does not prove that Ho is true. To prove with out any doubt that the null hypothesis is true, the population parameter would have to be known. This is usually not feasible.

The sample statistic is usually different from the hypothesized population parameter. For this reason we have to make a judgment about the difference.

If a hypothesized mean is 70 and the sample mean is 69.5 we musts make a judgment about the difference 0.5. Is it a true difference, i.e a significant difference, or is it due to chance / sampling. To answer this

question we conduct a test of significance, commonly referred to as a test of hypothesis.

***Identify the Alternative hypothesis ($H_1$)***: Alliterate hypothesis is a statement describes what we will believe if we reject the null hypothesis. It is designated $H_1$ (H sub – one) the alternate hypothesis will be accepted if the sample data provide us with evidence that the null hypothesis is false.

It is a statement that will be accepted if our sample data provide us with ample evidence that the null hypothesis is false.

***Step II: Determine the level of significance***

After setting up the null hypothesis and alternate hypothesis, the next step is to state the level of significance. It is the probability of rejecting the null hypothesis when it is actually true.

Level of significance is the risk we assume of rejecting the null hypothesis when it is a actually true.

The level of significance is designated by the Greek letter alpha, $\alpha$, it is also referred to as the level of risk.

Traditionally three levels of significance are known

0.05. level is selected for consumer research

0.01. for quality assurance

0.10. for political polling

The level of significance reflects the risk we want to assume A0.01 level of significance will yield smaller risk than 0.05 or 0.1.

The researcher must decide on the level of significance before formulating a decision rule and collecting sample data. This is very important to reduce bias. The level of significance can be any level between 0 and 1.

To illustrate how it is possible to reject a true hypothesis, suppose that a compute manufacturer purchase a component form a supplier. Suppose the contract specifies that the manufacture's quality assurance department will sample all incoming shipment of component. If more than 6% of the components sampled are substandard the shipment will be rejected.

The null hypothesis is:
$H_o$= the incoming shipment of components contains 6% or less substandard components.

The alternative hypothesis is:
$H_{1:}$ More than 6% of the components are defective.
A sample of 50 components just received revealed that 4 components or 8% were substandard.

The shipment was rejected because it exceeded maximum of 6%. If the shipment was actually substandard then the decision to return the component to the supplier was correct.

However suppose the 4 components selected in the sample were the only substandard components in the shipment of 4000 components. Only 1% were defective. In that case less than 6% of the entire shipment was substandard and rejecting the shipment was an error.

In terms of hypothesis testing we rejected the null hypothesis that the shipment was not subitandard when we should not have rejected it.

By rejecting a true hypothesis we committed a type I error.
A type I error is designated by $\alpha$ (alpha).

Type I error is rejecting the null hypothesis, Ho, when it is actually true.

The probability of committing another type of error, Type II error, is designated $\beta$, beta, failure to reject Ho when it is actually false.

The above firm would commit a type II error if, unknown to it, an incoming shipment contained 600 substandard components yet the shipment was accepted. Suppose 2 of the 50 component in the sample (4%) tested were substandard and 48 were good. Because the sample contains less than 6% substandard components, the shipment was accepted. But of all task the entire shipment 15% of the components we defective.

We often refer to those two possible errors as the alpha error $\alpha$, and the beta error $\beta$,

$\alpha$ error – the probability of making a type I error

$\beta$ error – the probability of making type II error

The following table shows the decision the researcher could make and the possible consequences.

| **Null Hypothesis** | **The researcher does not reject $H_o$** | **The Researcher rejects $H_o$** |
|---|---|---|
| If Ho is true | Correct decision | Type I error |
| If Ho is false | Type II error | Correct decision |

***Step III: Find the Test statistic***

Test statistic – A value, determined from sample information, used to reject or not to reject the null hypothesis.

There are many test statistics, Z (the normal distribution), the student t test, F, and $X^2$ or the chi –square.

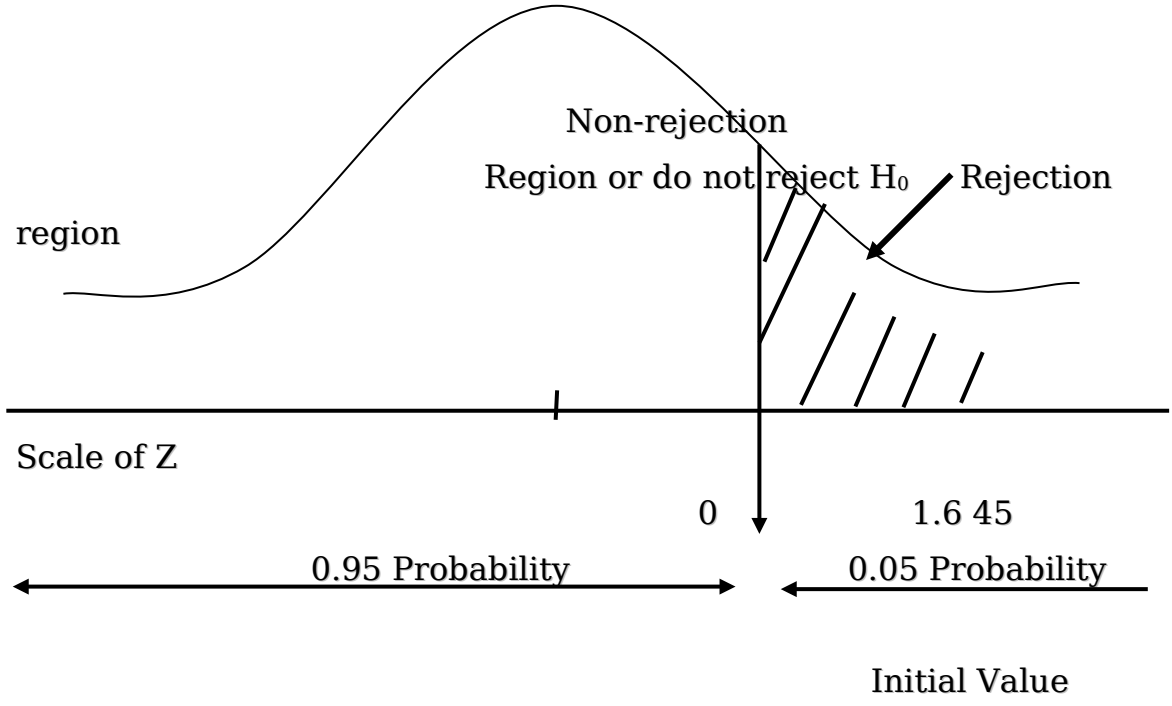
The standard normal deviate, Z distribution is used as test statistic when the sample size is large, n ≥ 30. Based on the sample size and the parameter to be tested the statistician will select the appropriate test statistic.

***Step IV: Determine the decision rule***

A decision rule is a statement of the conditions under which the null hypothesis is rejected and the conditions under which it is not rejected.

The region or area of rejection defines the location of all those values that are so large or so small that the probability of their occurrence under a true null hypothesis is rather remote.

Sampling distribution for the statistic Z, 0.05 level of significance.




The above chart portrays the rejection region for a test of significance. The level of significance selected is 0.05.

1. The area where the null hypothesis is not rejected includes the area to the left of 1.645

2. The area of rejection is to the right of 1.645
3. A one – tailed test is being applied /will be discussed latter on/
4. The 0.05 level of significant was chosen
5. The sampling distribution is for the test statistic Z , the standard normal deviate.
6. The value 1.645 separates the regions where the null hypothesis is rejected and where it is not rejected
7. The value 1.645 is called the critical value. It is the corresponding value of the test statistic for the selected level of significance i.e. Z value at the 0.05 level of significance is 1.645.

**Critical value:** The dividing point between the region where the null hypothesis is rejected and the region where it is not rejected.

***Steps V: Take a sample and made a decision***

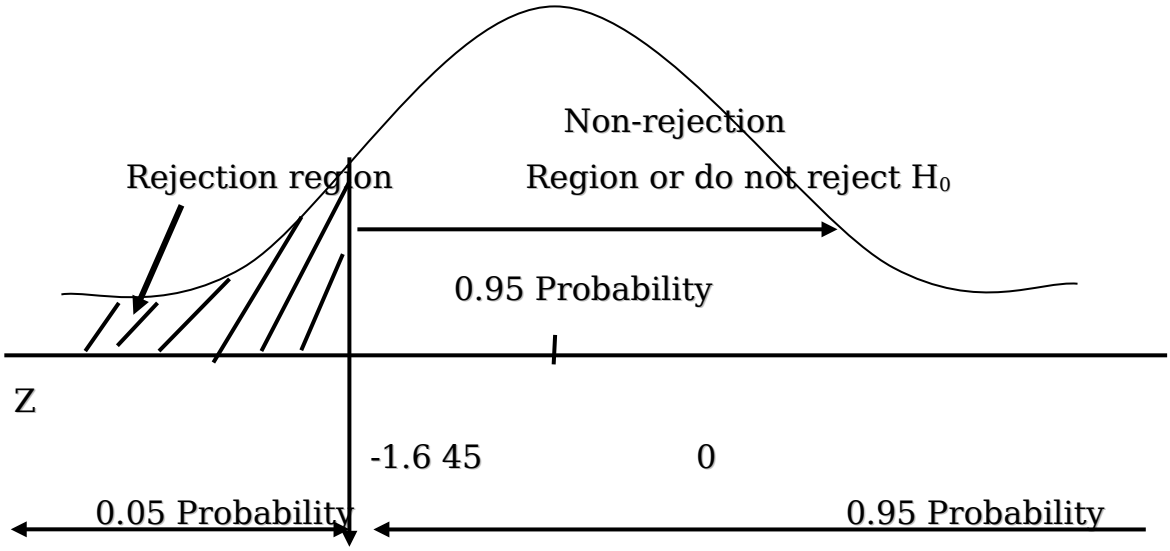
At this step a decision is made to reject or not to reject the null hypothesis. For the above chart, if based on sample data or information, Z is computed be 2.34 the null hypothesis is rejected at the 0.05 level of significance.

The decision to reject Ho is made because 2.34 lies in the region of rejection that is beyond 1.645. We would reject the null hypothesis reasoning that it is highly improbable that a computed Z value this large is due to sampling variation or chance. Had the computed value been 1.645 or less say 0.71 then Ho would not be rejected. It would be reasoned that such a small computed value could be attributed to chance that is sampling variation.

**One – Tailed and Two – Tailed tests of significance**

***One Tailed Test***

The region of rejection is only in one tail of the curve. The above example indicates that the region of rejection is in the right (upper) tail of the curve.




Initial Value

Consider companies purchase larger quantities of tyre. Suppose they want the tires to an average mileage of 40,000 Km of wear under normal usage. They will therefore reject a shipment of tires if accelerated - life test reveal that the life of the tires is significantly below 40000 Km on the average.

The purchasers gladly accept a shipment if the mean life is greater than 40000 Kms, they are not concerned with this possibility.

They are only concerned if they have sample evidence to conclude that the tires will average less than 40000 Kms of useful life.

Thus the test is set up to satisfy the concern of the companies that the mean life of the tires is less than 40000Km.

The null and alternate hypotheses are written: -

$H_o$: $\mu$ = 40,000 km and

$H_1$: $\mu$ < 40000 km

One way to determine the location of the rejection region is to look at the direction in which the inequality sign in the alternate hypothesis is pointing.

Test is one – tailed, if $H_1$ states $\mu$ > or $\mu$ < if $\mu_1$ , states a direction, test is one - tailed.

**Two-tailed test**

A test is two - tailed if $H_1$ does not state a direction.

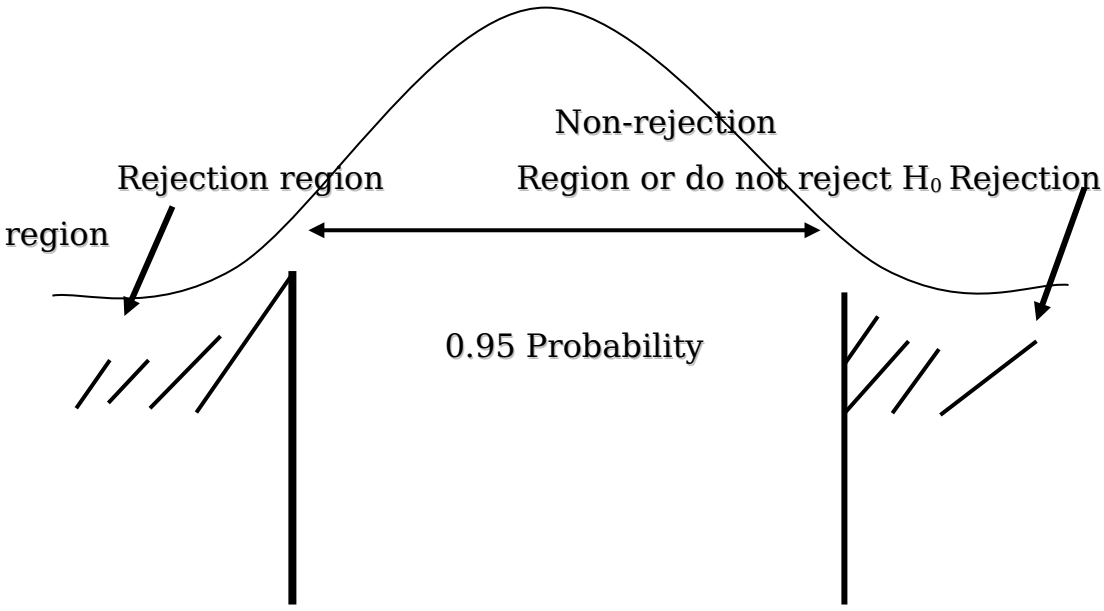
Consider the following example:

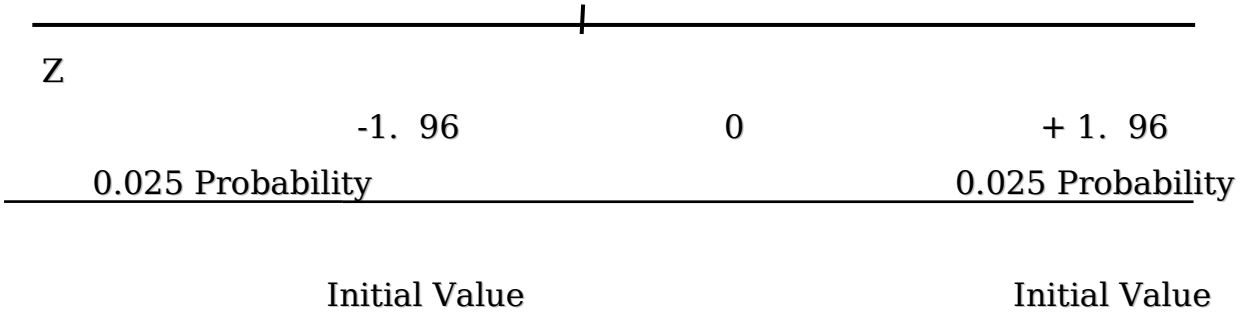
Ho: there is no difference between the mean income of males and the mean income of females.

$H_1$: there is a difference in the mean income of males and the mean income of females.

If Ho is rejected and $H_1$ accepted the mean income of males could be greater than that of females or vis versa. To accommodate these two possibilities, the 5 level of significance representing the area of rejection is divided equally in to two tails of the sampling distribution. If the level of significant is 0.05 each rejection region will have 0.025 probability.

Note that the total area under the normal curve is one found by 0.95 + 0.025 + 0.025.






## 5.4 HYPOTHESIS TESTING INVOLVING LARGE SAMPLE

Note that a sample of 30 or more is considered large.

### 5.4.1 Test for the Population Mean

#### 5.4.1.1 Population Standard Deviation Known

**Example.** The efficiency ratings of a company have been normally distributed over a period of many years. The arithmetic mean ($\mu$) of the distribution is 200 and the standard deviation is 19. Recently, however, young employees have been hired and new training and production methods introduced. Using the 0.01 level of significance, we want to test the hypothesis that the mean is still 200.

***Solution:***

**Step 1**. The null hypothesis is " The population mean is still 200 " the alternative hypothesis is “The mean is different from 200 " or "The mean is not 200"

the two hypotheses are written as:

Ho : $\mu$ =200

H1: $\mu \neq 200$

This is a two - tailed test because the alternate hypothesis does not state the direction of the difference.

That is, it does not state whether the mean is greater than or less than 200.

**Step 2:** - As noted the 0.01 level of significance is to be used. This is $\alpha$ the probability of committing a type I error. That is the probability of rejecting a true hypothesis.

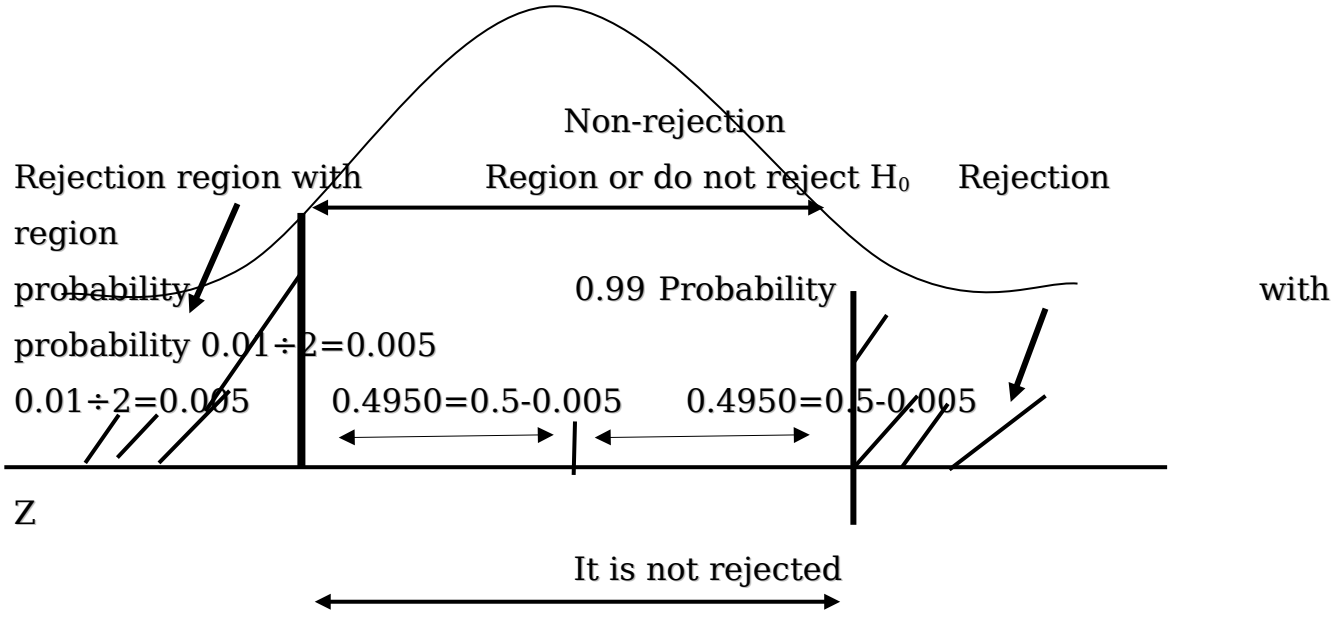
**Step 3:** - The test statistic for this type of problem is Z, the standard normal deviate /you will see later on that the sample size is large/

$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$$

**Step 4**: The decision rull is formulated by finding the critical values of Z from the table of normal distribution.

Since this is a two - tailed test, half of 0.01 or 0.005 is in each tail. Each rejection region will have a probability of 0.005.

The area where Ho is not rejected located between the two tails, is therefore, 0.99.
0.5000-0.005= 0.4950 so 0.4950 is the area between 0 and the critical value. The value nearest to 0.4950 is 0.495. The value for this probability is 2.58.



The decision rule is there fore: Reject the null hypothesis and accept the alternate hypothesis if the computed value of Z does not fall in the region between +2.58 and -2.58. Otherwise do not reject the null hypothesis.

**Step 5:** Take a sample and make a decision

Take a sample from the population (efficiently ratings) compute Z and based on the decision rule, arrive at a decision to reject Ho or not reject Ho.

The efficenty ratings of 100 employees were analyzed. The mean of the sample was computed to be 203.5.

Compute Z

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{203.5 - 200}{16/\sqrt{100}} = \frac{203.5 - 200}{1.6}$$ 203.5-200= 2.19

Since 2.19 does not fall in the rejection region, Ho is not rejected. So we conclude that the difference between 203.5, the sample mean, and 200 can be attributed to chance variation.

Note: Selecting the level of significance before setting up the decision rule and sampling the population is important not to be biased.

Ho is not rejected at the 1% level. We would have biased the later decision by not initially selecting the 0.01 level. Instead we could have waited until after the sampling and selected a level of significance that would cause the null hypothesis to be rejected. We could have chosen, for example , the 0.05 level. The critical value for that level are $\pm$ 1.96.

Since the computed value of Z (2.19) lies beyond 1.96 the null hypothesis would be rejected and we could concluded that the mean efficiency rating is not 200.

**Example 2:** The mean annual turn over rate of a brand of chemical is 6.0 (this indicates that the stock of the chemical turns over an average of six times a years) . The standard deviation is 0.5. It is suspected that the average turnover is not 6.0. The 0.05 level of significance is to be used to test this hypothesis.

1. State $H_o$, ad $H_1$
2. What is the value of $\alpha$?
3. Give the formula for the test statistic
4. State the decision rule
5. A random sample of 64 bottles of a brand was selected. The mean turn over rate computed to be 5.84. Shall we reject the null hypothesis at the 0.05 levels? Interpret.

**Solution:**

1. $H_o$: $\mu = 6.00$
   $H_1$: $\mu \neq 6.00$
2. 0.05
3. $Z = \dfrac{\overline{X} - \mu}{\sigma / \sqrt{n}}$
4. Do not reject the null hypothesis if the computed Z value fales between – 1.96 and + 1.96
5. $Z = \dfrac{5.84 - 6.00}{0.5 / \sqrt{64}} = 2.56$
6. reject Ho at the 0.05 level. Accept H1 the mean turnover is not equal to 6.00.

**A one Tailed Test**

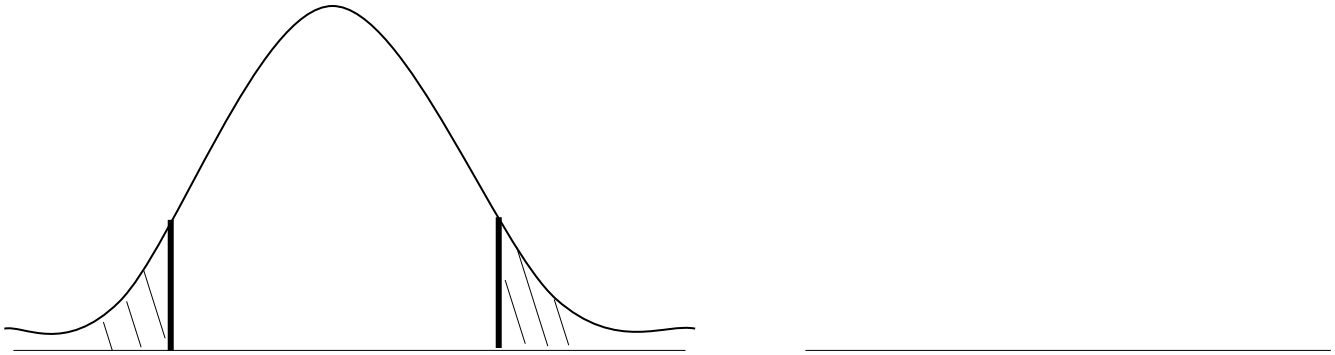
If the alternate hypothesis states a direction (either greater than “ or “ less than”) the test is one tailed. The hypothesis – testing procedure is

generally the same as for a two – tailed test, except that the critical value is different.

Let us change the alternate hypothesis in the previous problem, involving efficing racting of worker

$H_1$: $\mu \neq 200$ (tow – tailed test) to
$H_1$: $\mu > 200$ ( a one – tailed test )

The critical values for the two – tailed test were -2.58 and +2.58. The region of rejection for a one – tailed test is in the right tail of the curve
For a one-tailed test the critical value is found by
a. 0.5000 – 0.01 = 0.4900
b. The Z value for 0.4900 = probability is ± 2.33



## Check Your Progress –1

The management of chain of restaurants claims that the mean waiting time of customers for service is normally distributed with a mean of 3 minutes and a standard deviation of one minute. The quality assurance department found a sample of 50 customers at a restaurant and that the mean waiting time was 2.75 minutes. At the 0.05 significance level is the mean waiting time less than 3 minutes? (Note that this test is one tailed)

### P – values is Hypothesis Testing

Additional value is often reported on the strength of the rejection, or how confident we are in rejecting the null hypothesis. This method reports the probability (assuming that the null hypothesis is true) of getting a value of the test statistic at least as exterm as that obtained.

This procedure compares the probability, called P – Value, with the significance level.
If the P- value is smaller than the significance level, Ho is rejected. If it is larger than the significant level Ho is not rejected. This procedure not only results in decision regarding Ho but it gives us in sight into the strength of the decision.

A very small P- values say 0.001, means that there is a very little likelihood that Ho is true. On the other hand, a p- value of 0.4 means that Ho is not rejected, and we did not come very close to rejecting it.

Recall that for the efficiency ratings the computed value of Z was 2.19. The decision was not to reject Ho because the Z of 2.19 fall in the non-rejection area between 2.58 and + 2.58. The probability of obtaining a Z values of 2.19 or more is 0.0143 found by 0.5000 – 0.4857. To compute the P – value, we need to be concerned with values less than -2.19 and values greater than + 2.19. The p- value is 0.0286 found by 2(0.0143). The P – value of 0.0286 is greater than the significance level (0.01) decided upon initially, so Ho is not rejected.

## 5.4.1.2 Testing for the population mean: (standard deviation unknown)

In the preceding problems, we knew population standard deviation, $\sigma$. In most cases, however, it is unlikely that $\sigma$ would be known. Thus it must be

estimated using the sample standard deviation, S. Then the test statistic

$$Z = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

**Example:**

A department store issues it own credit card. The credit manger wants to find out if the mean monthly unpaid balance is more than Br. 400. The level of significance is set at 0.05. A random check of 172 unpaid balances revealed the sample mean to be 407 and the standard deviation of the sample 38. Should the credit manager conclude that the population mean is greater than 400, or is it reasonable to assume that the difference of 407- 400=7 is due to chance:

**Solution**

Ho: $\mu$ =400

Hi: $\mu$ > 400

Because $H_l$ states a direction, a one tailed test is applied. The critical value of Z is 1.645 for 0.05 level

$$Z = \frac{\overline{X} - \mu}{S/\sqrt{n}} = \frac{407 - 400}{380/\sqrt{172}} = 2.42$$

A value of this large (2.42) will occur less than 5% of the time. So the credit manager would reject the null hypothesis, Ho. that the mean unpaid balance is greater than 400, in favor of $H_1$, which states that the mean is greater than 400.

The P – value, in this one – tailed test is the probability that Z is greater than 2.42. Found by 0.5000-0.4922. 0.4922 is the probability that Z can assume a value of 2.420.

## Check Your Progress –2

At the time a server was heired at a restaurant was told by the manager that she can average more than 20 br a day in tips. Over the first 35 days

she was employed at the restaurant, the mean daily amount of her tips was 24.85 br with a standard deviation of 3.24 br. At the 0.01 significance level, can the manager conclude that she is earning more than 20 br. per day in tips?

### 5.4.2 Hypothesis testing; Two-population means; Independent population

**Assumption for two-sample test**

1. The population should be normally distributed
2. The population standard deviations for both population should be known. If they are not known, then both samples should contain at least 30 observations so that the sample standard deviation can be used to approximate the population standard deviation
3. The samples should be drawn from independent population.

If we select random samples from two normal population the distribution of the differences between the two means is also normal or if a large number of independent random samples are selected from two population, the difference between the two means will be normally distributed. If these differences are divided by the standard error of the difference, the result is the standard normal distribution.

The formula for the test statistic Z is

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

$\bar{X}_1 - \bar{X}_2$ ← The difference between two sample means

$\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$ ← Standard error of the difference between two

**Example:** Each patient at a hospital is asked to evaluate the service at the time of discharge. Recently there have been several complaints that resident physicians and nurses on the surgical wing respond too slowly to the emergency calls of senior citizens. The administrator of the hospital asked the quality assurance department to investigate. After studying the

problem, the quality assurance department collected the following sample information. At the 0.01 significance level, is the response time longer for the senior citizens, emergencies?

| **Patient type** | **Smaple mean** | **Sample standard deviation** | **Sample Size** |
|---|---|---|---|
| Senor Citizens | 5.5 Minutes | 0.40 minuets | 50 |
| Other | 5.3 Minutes | 0.30 minutes | 100 |

***Solution:-***

The testing procedure is the same as for one sample test except the formula for the test statistic, Z:

Step 1: Ho: there is no difference in the mean response time between the two groups of patients.

i: e The difference of 0.2 minute, in the arithmetic mean response time is due to chances.

$H_1$: the mean response time is greater for the senior citizens

Because the quality assurance department is concerned that the response time is greater for senior citizens, he wants to conduct a one – tailed test. There fore the null and alternate hypotheses are stated as follows.

$H_o$: $\mu_1 = \mu_2$

$H_1$: $\mu_1 > \mu_2$

Step 2: The 0.01 significance level is selected.

Step 3: the test statistics is Z, the standard normal distribution, $Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$

Step 4: The decision rule is:

Reject the null hypothesis if the computed value of Z is greater then 2.33. The critical value for 0.01 cruel, one-tailed test is 2.33

Step 5: Calculate the test statistic and make a decision.

The test statistic is $Z = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}$

$$Z = \frac{5.5 - 5.3}{\sqrt{\dfrac{(0.40)^2}{50} + \dfrac{(0.30)^2}{100}}} = 3.13$$

The computed value of 3.13 is beyond the critical value of 2:33. Therefore, the null hypothesis is rejected and the alternate hypothesis is accepted at the 0.01 significant level.

The quality assurance department will report to the administrator that the mean response time of the nurses and resident physicians is longer for senior citizens than for other patients.

What is the P-value in this problem?

P- Value is the probability of computing aZ value this large or larger when Ho is true.

What is the likelihood of aZ value greater than 3.13

P(Z=3.13)= 0.4991

So, P(Z)>31.13 ) =0.5000-0.44991=0.0009

Ho is very likely false and there is little likelihood of a type I error.

## Check Your Progress –3

A peal Estate Association is preparing a pamphlet that they feel might be of interest to prospective home buyers in the eastern and western areas of the city. One item of interest is the length of time the seller occupied the home. A sample of 40 home sold recently in the eastern areas

revealed that the mean length of ownership was 7.6 years with standard deviation of 2.3 years.

A sample of 55 homes in the western areas reaealled that the mean length of ownership was 8.1 years with a standard deviation of 2.9 years. At the 0.05 significance level can we conclude that the Eastern residents owned the homes for a shorter period of time?

### 5.4.3 Testing for Population Proportion

In testing hypothesis for the population proportion the assumptions of the binomial distribution should be met. To test for the proportion

a) np and n(1-p) both should be greater than 5.
b) n should be at least 50

**Example:** suppose prior elections in a region indicated that it is necessary for a candidate for governor to receive at least 80% of the majority vote. The incumbent governor is interested in assessing his chance of returning to office and plans to have a survey conducted consisting of 2000 registered voters

Using the five – step hypothesis testing procedure, asses the governor's chances of reflection

np = 2000(0.8) = 1600 which is greater than 5
nq = n(1-p) = 2000(1-0.8) = 400 which is greater than 5
both 1600 and 400 are greater than 5

Step 1: The null hypothesis Ho is that the population proportions is 0.80
The alternate hypothesis, $H_1$ is that the proportion is less than 0.80.
The incumbent governor is concerned only when the sample proportion is less than 0.8. If it is equal to or greater than 0.8 he will have no problem; that is the sample data would indicate he will be probably be reelected.

Ho: P = 0.80

$H_1$: P<0.80

Step 2: The level of significance is 0.05

Step 3: Z is the appropriate statistic

$Z = \frac{\bar{P} - P}{\sigma_p}$ where P – is the population proportion and

$\bar{P}$ is the sample proportion, $\sigma_p$ is the standard error of the proportion

$\sigma_{\bar{P}} = \sqrt{\frac{\bar{p}(1 - \bar{p})}{n}}$ so the formula for Z becomes :

$$Z = \frac{\bar{p} - p}{\sqrt{\frac{P(1 - P)}{n}}}$$

**Step 4:**

The area between 0 and the critical value is, 1.645 obtained for the Z table 0.45000 = 0.5000 – 0.05 Z value for probability 0.450 is 1.645.

The decision rule is therefore reject the null hypothesis and accept the alternate hypothesis if the computed value of Z falls to the left of -1.645 otherwise do not reject Ho.

**Step 5.** Take Sample and make a decision with respect to Ho.

The sample survey of 2000 potential voters revealed that 1550 planned to vote for the incumbent governor. Is the proportion of 0.775 (found by 1550/2000) close enough to 0.80 to conclude that the difference if due to chance?

n =2000

$\bar{P} = \frac{1550}{2000} = 0.775$

p = 0.80, the hypothesized population proportion

$Z = \frac{\bar{P} - P}{\sqrt{P(1 - P)/n}} = \frac{0.775 - 0.80}{\sqrt{\frac{0.8(1 - 0.801}{2000}}} = -2.80$

The computed value of Z (-2.80) is in the rejection region. So the null hypothesis is rejected at the 0.05 level of significance. The difference of 2.5 percentage points between the sample (77.5) and the hypothesized population percentage (80.0) is statistically significance. It is probably not due to sampling variation.

To put it another way the evidence at this point does not support the claim that the incumbent governor will return to the office.

The p- Value is 0.0026 found by 0.5000-0.4974. 0.4974 is the probability of Z to assume
–2.80 value. It is less than the significance level of 0.05. So Ho should be rejected. This further indicates that the likelihood that Ho is ture is small.

## Check Your Progress –4

This Claim is to be investigated at the 0.02 level “Forty percent of those persons who retired from an industrial job before the age of 60 would return to work if a suitable job were available” 74 persons out of the 200 sampled said they would return to work.

Can we conclude that the fraction returning to work is different from 0.40?

1) Can the Z test be used? Why or why not?
2) State the null hypothesis and the alternate hypothesis
3) Compute Z, and arrive at a decision

## 5.4.4 Testing for the Difference between two Population Proportions

**Example:** - a company has developed a new perfume
One of the questions is whether the perfume is preferred by a larger proportion of younger women or a larger proportions of older women. A standard smell test is used.

Women selected at random are asked to sniff several perfumes in succession, including the new. Each woman selects the perfume she likes best.

**Step 1**

Ho " There is no difference between the proportion of younger women who prefer the perfume and the proportion of older women who prefer it"

If the proportion of younger women in the population is designated as $P_1$ and the proportion of older women is $P_2$ then;

Ho: $P_1 = P_2$

The alternate hypothesis is that the two proportions are not equal or:

Hi: $P_1 \neq P_2$

**Step 2:** It was decided to use the 0.05 level.

**Step 3:** The test statistic is Z and the formula is: -

$$Z = \frac{\bar{P} - \bar{P}_2}{\sqrt{\frac{\bar{P}_c(1 - \bar{P}_c)}{n_1} + \frac{\bar{P}_c(1 - \bar{P}_c)}{n_2}}}$$

where: $n_1$ , is the number of young women selected in the sample $n_2$ is the number of older women selected in the sample, $\bar{P}_c$ = is the weighted mean of the two sample proportion computed by

$$\bar{P}_c = \frac{\text{Total number of success}}{\text{Total number of sample}} = \frac{x_1 + x_2}{n_1 + n_2}$$

where $x_1$ is the number of younger women (sample 1) who prefer the perfume, $x_2$ is the number of older women (sample 2) who prefer the perfume.

$\bar{P}_c$ is generally referred to as the pooled estimate of the population proportion or it is a combined estimate, combined proportion.

**Step 4: The Formulate Decision Rule:**

The critical values for the 0.05 level two-tailed tests are -1.96 and +1.96. If the computed Z value is in the region between +1.96 and -1.96, the null hypothesis will not be rejected. If it does occur it is assumed that any difference between the two proportions is due to chance variation.

Two – tailed test, Areas of rejection and Non-rejection 0.05 level of significance.

**Step 5: The decision**

A total of 100 young women selected at random, and each was given the standard smell test. Forty of the 100 young women chose the perfume, as they liked best

$x_1 = 40$

$n_1 = 100$ and

200 older women were selected at random and each was given the same standard smell test of the 200 women 100 preferred the perfume.

$x_2 = 100$

$n_2 = 200$

$$\bar{P}_1 = \frac{x_1}{n_1} = \frac{40}{100} = 0.40$$

$$\bar{P}_2 = \frac{x_2}{n_2} = \frac{100}{200} = 0.50$$

The pooled or weighted proportion $\bar{P}_c$ is

$$\bar{P}_c = \frac{x_1 + x_2}{n_1 + n_2} = \frac{40 + 100}{100 + 200} = 140 / 300 = 0.4667$$

$$\mathbf{Z} = \frac{\bar{P}_1 - \bar{P}_2}{\sqrt{\frac{\bar{P}_c(1 - \bar{P}_c)}{n_1} + \frac{\bar{P}_c(1 - \bar{P}_c)}{n_2}}} = \frac{40 - 0.50}{\sqrt{\frac{0.4667(0.5333)}{100} + \frac{0.4667 + 0.5333}{200}}} = -1.64$$

The computed value of Z (-1.64) falls in the non-rejection region. Therefore we concluded that there is no difference in the proportion of younger and older women who prefer the perfume. In this case we expect the P- value to be greater than the significance level of 0.05, and it is.

for Z = -1.64 probability is 0.4495

P value = 0.5000 – 0.4495 = 0.0505 for one tail only

However the test was two tailed, so we must account for the area beyond 1.64 as well as the area less than -1.64. Then
The P – value is 2(0.0505) = 0.1010

### Check Your Progress –5

Of 150 girls who tried a new candy 87 rated it excellent of 200 boys sampled 123 rated it excellent using the 0.10 level of significance, can we conclude that there is a difference in the proportion of girls versus boys who rate the candy excellent?

1. State the null and alternate hypotheses
2. What is the decision rule
3. Compute the value of the test statistics
4. State your decision granting Ho
5. Compute the P – value

## 5.5 STUDENT'S t TEST/ SMALL SAMPLE/

When the population is normal and the standard deviation is known the Z distribution is employed as a test statistic for a test. If the population standard deviation is not know the sample standard deviation is substituted for $\sigma$. If the sample size is at least 30, the results are deemed satisfactory.

If the sample size is less than 30 observations and $\sigma$ is unknown the Z distribution is not appropriate. The student's t or the t distribution is used as the test statistic.

### 5.5.1 Characteristics of Student's t Distribution

**Note:** The Characteristics of student's distribution are discussed in unit 4. To mention some

1. It is a continuous distribution.
2. It is bell- shaped and symmetrical,
3. There is not one distribution, but rather a “family” of t distribution. All have the
   small mean of zero but their standard deviations differ according to the sample size n. The t distribution for a sample size of 20,22, 25 are different.
4. It is more spread out and flat at the center than is the Z. However as the sample size increases, the curve representing the t distribution approaches the Z distribution. If the sample size is 30 we will have approximately the same t distribution as the Z.

Since the t distribution has a greater spread or the tails are wide, the critical values of t for a given level of significance are larger in magnitude than the corresponding Z critical values.

Region of rejection for the Z and t distribution 0.05 level, one tailed test

Why the critical value for a given level of significance is greater for small samples than for large samples?

a. The confidence interval will be wider than for large samples using the Z distribution
b. The region where Ho is not rejected is wider than for large samples using Z distribution
c. A larger t value will be needed to reject the null hypothesis than for large samples using Z. In other words because there is more variability in sample means computed from smaller samples we are less apt to reject the null hypothesis.

### 5.5.2 A Test for the Population Mean

**Example**: Experience in investigating accident claims by an insurance company revealed that it cost 60 on the average to handle the paper work, pay the investigator, and make a decision. The cost compared with that of other insurance firms was deemed exorbitant, and cost cutting measures were instituted. In order to evaluate the impact of these new measures, a sample of 26 recent claims was selected at random and cost studies were made. It was found that the sample mean, $\bar{x}$, and the sample standard deviations, s, were 57 and 10 respectively.

At the 0.01 level is there a reduction in the average cost, or can the difference of 3 = (60-57) be attributed to chance?

The usual five-step hypothesis testing procedure is used

**Step 1:** - the null hypothesis, Ho: the population mean is 60

The alternate hypothesis, $H_1$ the population mean is less than 60. i.e.

Ho: $\mu = 60$

H1:- $\mu < 60$

**Step 2:** The 0.01 level is to be used

**Step: 3** the test statistic is student's t distribution. Because the population standard deviation is unknown and the sample size is small (26 under 30)

$$t = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

**Step 4:** The critical value of t are given in table 4

There are n -1 degrees of freedom for the test df (26-1= 25)

The critical value for df = 25, a one tailed test and 0.01 level is 2.485

The decision rule for this one tailed test is reject Ho if the computed value of t falls in any part of the tails to the left of –2.485 otherwise do not reject Ho.

Ho: N= 60

Ho; $\mu = 60$

H1: $\mu < 60$

df = 26 – 1 = 25

**Step 5:** Compute t, and arrive at a decision

$$t = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

$\overline{X} = 57$

$\mu = 67$

$$t = \frac{57 - 60}{10/\sqrt{26}} = -1.530$$

S = 10

n = 26

Because -1.530 lies in the region to the right of the critical value –2.485 Ho is not rejected at the 0.01 level.

This indicates that the cost cutting measures have not reduced the mean cost per claim to less than 60 based on sample results.

## Check Your Progress –6

From past records it is known that the arithmetic mean life of a battery used in a digital clock is 305 days. The lives of the batteries is normally distributed. The battery was recently modified to last longer. A sample of 20 modified batteries were tested. It was discovered that the man life was 311 days and the sample standard deviation was 12 days. At the 0.05 level of significance, did the modification increases the mean life of the battery?

1. State the null and alternate hypotheses
2. State the decision
3. Compute t and make a decision

### 5.5.3 Comparing two Population Mean

A test using the t distribution can also be applied to compare two sample means to determine if the samples were obtained from normal populations with the same mean.

Three assumption are required to test for two population means.

1. The populations must be normally distributed (or approximately normally distributed)
2. The populations must be independent
3. The population variance must be equal

The statistic for the two sample is similar to that employed for the Z statistic except that an additional calculation is required.

The two-sample variance must be polled to form a single estimate of the unknown population variance. Since the samples have fewer than 30 observations the population standard deviations, are not known. So, we substitute $S^2$ for $\sigma^2$, because we assume that the two populations have equal variances, the best estimate we can make of that value is to combine or pool all the information we have with respect to the population variance.

The following formula is used to pool the sample variances. Notice that two factors make up the weights: - the number of observations in each sample and the sample variances themselves. The pooled variance, $_{Sp}^2$ is

$$_{Sp}^2 = \frac{(n-1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

where $S_1^2$ – variance of sample one

$S_2^2$ – variance of sample two and

$n_1 + n_2 - 2$ is total df.

The value of t is then determined by the formula

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where: $\overline{X}_1$ is sample mean one

$\overline{X}_2$ is sample mean two

$n_1$ is sample size for first sample

$n_2$ is sample size for second sample

The number of degrees of freedom in the test is equal to the total number of items sampled minus the number of sample. Since there are two samples, there are

$n_1 + n_2 - 2$ degrees of freedom.

**Example:** Two different procedures are proposed for mounting engine on a frame. The question is: 'is there a difference in the mean time to mount the engine on the frame?' To evaluate the two proposed methods, it was decided to conduct a time and motion study. A sample of five employees were timed using procedure 1 and 6 were timed using procedures 2.

The results in minutes, are:

| Procedure 1 (Minutes) | Procedure 2 ( Minutes ) |
|---|---|
| 3 | 2 |
| 7 | 4 |
| 5 | 9 |
| 8 | 3 |
| 4 | 2 |
| 3 | |

Is there a difference is the mean mounting times? Use the 0.10 significance level.

**Solution :**

The null hypothesis states that there is no difference in mean mounting time between the two procedures and the alternate hypothesis states that there in a difference is the mean mounting time between the two procedures.

**Step I.** Ho: $\mu_1 = \mu_2$ $H_1$: $\mu_1 \neq \mu_2$

The required assumptions are met.

The degrees of freedom are determined by $n_1 + n_2 - 2$ there are 9 degrees of freedom

(5 + 6-2).

**Step II.** The 0.01 level is to be used

**Step III.** The test statistic is $t = \dfrac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$

**Step IV**. The critical value of t for df = 9, a two tailed test, at the 0.10 level of significance, are + 1.833 and -1.833

We do not reject the null hypothesis if the computed t value falls between -1.853 and +1.833 otherwise Ho is rejected.

**Calculate t and make the decision**

(a) Calculate the sample variance

| Procedure 1 | | Procedure 2 | |
|---|---|---|---|
| $X_1$ | $X_1^2$ | $X_2$ | $X_2^2$ |
| 2 | 4 | 3 | 9 |
| 4 | 16 | 7 | 49 |
| 9 | 81 | 5 | 25 |
| 3 | 9 | 8 | 64 |
| 2 | 4 | 4 | 16 |
| $\Sigma x_1 = 20$ | $\Sigma x_1^2$ 114 | 3 | 9 |
| | | $\Sigma x_1 = 30$ | $\Sigma x_2^2 = 172$ |

$$S_1^2 = \frac{\sum X_1^2 - \frac{(\sum X_1)^2}{n_1}}{n - 1} \qquad S_2^1 = \frac{\sum X_2^2 - \frac{(\sum X_2)^2}{n_2}}{n_2 - 1}$$

$= \frac{144 - \frac{(20)^2}{5}}{5-1} = 8.5$ $= \frac{172 - \frac{(30)^2}{6}}{6-1} = 4.44$

(b) Pool the variances

$$S_p^2 = \frac{(n-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} = \frac{(5-1)(8.5)+(6-1)(4.4)}{5+6-2} = 6.222$$

(c) Determine t

$\overline{X}_1 = 20/5 = 4$ and $\overline{X}_2 = \frac{30}{6} = 5$

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{Sp^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{4-5}{\sqrt{6.222\left(\frac{1}{5}+\frac{1}{6}\right)}} = -0.6626$$

The decision is not to reject Ho because -0.6620 falls in the region between -1.833 and + 1.833.We conclude that there is no difference in the mean time to mount the engine on the frame.

**Check Your Progress –7**

The net weight of sample of bottles filled by two different machines produced by two different manufactures, are ( in grams )

Machine 1-5,8,7,6,9,7

Machnies 2-8,10,11,9,12,14,9

At the 0.05 level is the mean might of the bottled filed by machine 2 are greater than the mean weight of the bottles filled by machine 1? (Note that the test is one tailed)

## 5.5.4 Hypothesis Testing Involving Paired Observations

There are situations where the samples are not independent. A particular group will be exposed to two different experiments. In a sense the sample is one.

**Example:** The production manager wants to find out whether a unique training program will increase employee efficiency.

He plans to take a random sample of 10 employees and record their efficiency before the training starts. After completion of the program, the efficiency of the same sample of employees will be recorded.

Thus there will be a pair of efficiency ratings for each member of the sample. A test of hypothesis is conducted to find out if there is a difference between the ratings before and after the training program. It is called a paired difference test

The sample dates are

| Sample member | Efficiency Ratings | | Difference (d) | difference /// squared $d^2$ |
|---|---|---|---|---|
| | Before | After | | |
| 1 | 128 | 135 | 7 | 49 |
| 2 | 105 | 110 | 5 | 25 |
| 3 | 119 | 131 | 12 | 144 |
| 4 | 140 | 142 | 2 | 4 |
| 5 | 98 | 105 | 7 | 49 |
| 6 | 123 | 130 | 7 | 49 |
| 7 | 127 | 131 | 4 | 16 |
| 8 | 115 | 110 | -5 | 25 |
| 9 | 122 | 125 | 3 | 9 |
| 10 | 145 | 149 | 4 | 16 |
| | | | | $\sum d = 46$ |

$$\sum d^2 = 386$$

For the test of hypothesis to be conducted, there is essentially only one sample, not two. We are testing the hypothesis that the distribution of the differences has a mean of 0.

The sample is made up of the differences b/n the efficiency ratings before the training program and the ratings after the program.

If production methods before and after the training program remain the same, one could logically expect some employees to benefit from the training program and to become more efficient. Other employees would prefer the method used before the training program,. And their efficiency would remain the same or even decrease. Thus the mean of the difference in efficiency ratings designated $\mu_d$ would balance out and equal zero.

The production manager wants to know whether or not the new production technique affect efficiency. If it does one would reasonably assume that most of the difference would be positive i.e. increased efficiency.

The null hypothesis to be tested is therefore; the mean difference is zero or there is no difference in the efficiency ratings before and after the training.

Ho: $\mu d = 0$.
The alternate hypothesis is that the mean of the difference is greater than O
$H_1$: $\mu_d > 0$, signifying that the differences are positive.
The test statistic t is

$$t = \frac{\bar{d}}{S_d / \sqrt{n}}$$

where $\bar{d}$ = the mean difference i.e., $\frac{\sum d}{n}$

$S_d$ = standard deviation of the differences between the paired observations

The standard deviation of the differences is computed as

$$S_d = \sqrt{\frac{\sum d^2 - \frac{(\sum d)^2}{n}}{n-1}}$$

The critical value of t for this one tailed test of paired difference for 9 degree of freedom at the 0.05 level is 1.833

$$\bar{d} = \frac{\sum d}{n} = \frac{46}{10} = 4.60$$

$$S_d = \sqrt{\frac{\sum d^2 - \frac{(\sum d)^2}{n}}{n-1}} = \sqrt{\frac{386 - \frac{(46)^2}{10}}{10-1}} = 4.40$$

$$t = \frac{\bar{d}}{Sd/\sqrt{n}} = \frac{4.6}{4.4/\sqrt{10}} = 3.33$$

Because the value of t (3.30) lies in the rejection rejoin, that is beyond the critical value of 1.833, the null hypothesis is rejected.

The production manger has convincing evidence that this special training program will be effective in increasing efficiency.

## Check Your Progress –8

An Agricultural Experimental Station plans to test the effectiveness of two solutions for corn seeds to increases resistance for a particular type of pest and increase germination and growth times. The purpose of the experiment is to determine if there is a difference in effectiveness of two solutions, solution A and solution B.

Various corn seeds are to be used in the experiment. A pair of seeds is selected one is soaked in solution A, the other in solution B. Then they are planted and the germination and growth times in days are recorded.

| | Pair | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Solution | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| A | 16 | 9 | 21 | 14 | 26 | 27 | 18 | 14 | 30 |
| B | 18 | 7 | 26 | 11 | 26 | 27 | 19 | 20 | 28 |

1. State the null and alternative hypothesis.
2. Using the 0.05 level what is the critical value?
3. Using the above nine pairs of sample compute t and arrival at a decision.

## 5.6 TESTING FOR DIFFERENCES OF VARIANCES / THE 'F' DISTRIBUTION FOR COMPARING TWO POPULATION VARIANCES

Determining whether or not one normal population has more variation than the other is important for many decision-making purposes in business.

Suppose two machines are set to produce steel bars of the same length. The bars, therefore, should have the same mean length. We want to ensure that, in addition to having the same mean length, they have similar variation. or

The mean rate of return on investment of two types of projects may be the same. But there may be more variation in the return of one than the other. Decision, as to which project is more feasible, is based on the level of variation.

The F distribution is used to test the hypothesis that the variation of one normally distributed population equals the variance of another normally distributed population.

The major characteristics of the F distribution:

a. There is A "Family" of F distribution. A particular member of the family is determined by two parameters, the degree of freedom in the numerator, and the degree of freedom in the denominator. There is only one F distribution for the combination of 29 degree of freedom in the numerator and 28 degrees of freedom in the denominator. The shape of the curves changes as the degrees of freedom change.
b. F cannot be negative; and it is a continuous distribution.
c. The curve representing an F distribution is positively skewed. F can not be negative.

d. Its value range from 0 to positive infinite (∞. As the value of F increases the curve approaches the x-axis, but it never touches it.

For all investigations the null hypothesis is that 'the variance of one normal population $\sigma_1^2$, equals the variance of the other normal population $\sigma_2^2$.' To conduct the test, a random sample of n, observations is obtained from one population and a sample of $n_2$ observations is obtained from the second population. The test statistics is

$S_1^2 / S_2^2$ , where $S_1^2$ and $S_2^2$ are the respective sample variances.

If the null hypothesis is true Ho: $\sigma_1^2 = \sigma_2^2$.

The test statistic follows the F distribution with $n_1 - 1$ and $n_2 - n_1$ degrees of freedom

The larger sample variance is placed in the numerator; hence, the F ratio is always positive and greater than one. Thus, the upper-tail critical value is the only one required. The critical value of F is found by dividing the significance level in half $(\alpha/2)$ and then referring to the appropriate number of degrees of freedom in the F table.

**Example:** A car rental offers limousine service from city center to the airport. The manager of the company is considering two routes. He wants to conduct a study of both routs and then compare the results. He recorded the following data. Using the 0.10 significance level, is there a difference in the variation in the two routes?

| Route Sample | Mean Time (minutes) | Standard deviation (Minutes) | size |
|---|---|---|---|
| 1 | 56 | 12 | 7 |
| 2 | 59 | 5 | 8 |

The manager noted that the mean times seem very similar but there is more variation, as measured by the standard deviation, in route 1,

The reason can be route 1 contains more stoplights, while the distance is shorter but for rout 2 the distance is longer but it is a limited access high way. So he decides to conduct a statistical test to determine if there is really a difference in the variation of the two routes.

The usual five-step hypothesis testing procedure will be employed.

**Step 1:** the test is two-tailed because we are looking for a difference in the variation of the two routes. We are not trying to show that one route has more variation than the other.

Ho: $\sigma_1^2 = \sigma_2^2$

$H_1$: $\sigma_1^2 \neq \sigma_2^2$

**Step 2:** A significant level of 0.10 is selected.

**Step 3:** the appropriate test statistic is F distribution $F = \frac{S_1^2}{S_2^2}$

**Step 4:** The decision rule is obtained from the F table because we are using a two tailed test the significance level is 0.05 found by $\alpha/2 = \frac{0.10}{2}$ there are $n_1 - 1 = 7\text{-}1 = 6$ degrees of freedom in the numerator and $n_2 - 1 = 8 - 1 = 7$ degrees of freedom in he denominator. The critical value for a 0.05 level and df(7,6) is 3.87. If the ratio of the sample variances $S_1^2 / S_2^2$ exceed 3.87 the null hypothesis is rejected.

**Step 5:** the computed value of the tests statistic is 5.70 = $\frac{S_1^2}{S_2^2} = \frac{(12)^2}{(5)^2}$

The null hypothesis is rejected and the alternate hypothesis accepted. The variation is not the same in the two pouts.

The usual procedure is to determine the F ratio by putting the larger variance in the numerator. This will force the F ratio to be larger than 1.0. Why is this necessary?

It allows us to always use the upper tail of the F statistic thus avoiding the need for more extensive F tables.

How a one-failed testis to be handled: Again we will arrange the F ratio so that it is always greater that 1.00. Under these conditions it is not necessary to divide the level of significance in half. We are there fore restricted to the 0.05 of 0.1 level and significance for one-tailed tests in the F table.

## Check Your Progress –9

A company assembles electrical components. For the last 10 days employee ‘A’ averaged 9 rejects per day with a standard deviation of 2 rejects. Employee ‘B’ averaged 8.5 rejects per day with a standard deviation of 1.5 rejects over that same period. At the 0.05 level, can we

conclude that there is more variation in the number of rejects per day attributed to employed A? (Note that the givens are standard deviations not variances. The test is one-tailed)

## 5.7 ANSWERS TO CHECK YOUR PROGRESS

1. Reject Ho  Z = 1.767 > 1.645
2. Reject Ho;  Z = 8.86 > 2.33
3. Do not reject Ho;  Z = 0.936 > 1.645
4. 1) Yes, because Np and n(1 – p) or nq exceeds
   2) Ho: p = 0.40
      p ≠ 0.40
   3) Do not reject Ho  Z = -0.866 > -2.58
5. 1) Ho: $P_1 = P_2$  H: $P_1 \neq P_2$
   2) Reject Ho if Z < -1.645 or > 1.645
   3) $\bar{P}c = \frac{87+123}{150+200} = 0.6$

      $Z = \frac{0.55 - 0.615}{\sqrt{\frac{0.6(0.4)}{150} + \frac{0.6(0.4)}{200}}}$

   4) Do not reject Ho
   5) P-value = 2(0.5000 – 0.2454) = 0.5092
6. 1) Ho: $\mu = 305$  $H_1$: $\mu > 305$
   2) Reject Ho if t > 1.729
   3) t = 2.236
   reject Ho, the mean is greater than 305 age
7. 1) Ho: $\mu_1 = \mu_2$  $H_1$  $\mu_1 < \mu_2$
      Reject Ho if t < -1.782
      t = -2.827
      reject Ho
8. Ho: $\mu_d = 0$  $H_1$  $\mu_d \neq 0$
   critical values are –2.306 and +2.306

$t = \frac{\bar{d}}{Sd/\sqrt{n}} = \frac{0.22}{3.667/\sqrt{9}} = 0.180$

Do not reject Ho. There is no difference between the two solutions

8. Critical value is 3.18

   If F > 3.18 reject Ho

   $F = \frac{(2)^2}{(1.5)^2} = 1.78$

   Do not reject Ho

## 5.8 MODEL EXAMINATION QUESTIONS

1. **Test for the population mean**

   An educator claims that the average IQ of city collage students is no more than 110. To test this claim, a random sample of 150 was taken and given relevant tests. Their average IQ score came to 111.2 with a standard deviation of 7.2. At level of significance of 0.01, test if the claim of the educator is justified.

2. **Test for two population means**

   A potential buyer of electric bulbs bought 100 bulbs each of two famous brands, A and B. Upon testing both these samples, he found that brand A had a mean life of 1500 hours with standard deviation of 50 hours whereas brand B had an average life of 1530 hours with a standard deviation of 60 hours. Can it be concluded at 5% level of significance that the two brands differ significantly in quality?

3. **Test for the population proportion**

   A sociologist taken a survey of the previous lottery winners of one million br. has taken and found that 80% of these winners continue to work on their job. A psychologist felt otherwise. To test the report of the state, he took a sample of 100 such winners at random and found that only 25 winners of this sample had quit their jobs. At

95% confidence level, can we conclude that the state report is correct?

4. **Test for the difference between two population proportions**

   Random samples of 2000 people in town A and 3000 in town B were asked if they thought there was too much violence on TV these days. 1400 people in town A and 1800 people in town B replied in the affirmative. Can we conclude at 99% confidence level that proportions are significantly different?

5. **Test for the population mean**

   A Home Owners' Association has determined that the average number of days a house was in the market for sale was 90 days, before it was sold. A real estate agency believes that in certain section of Long Island, the average number of days the houses remained in the market before sales was less than 90. It selected a random sample of 10 homes that were sold in this section in order to justify what it believes. The following data represents the number of days that each of these 10 homes stayed in the market before sale?

   87, 95, 78, 83, 110, 75, 82, 92, 90, 80

   At 0.01 level of confidence and assuming that population is approximately normal, is the real estate agency justified in its belief?

6. Test for the two population means

   Two drug manufacturing companies produce headache remedies. Each company claims that its drug bring faster-acting relief. A consumer protection agency wants to test if one drug brings relief faster than the other. An experiment was performed to compare the mean lengths of time required for bodily absorption of both drugs. 12 people selected at random were given dosage of one drug and

another 12 people randomly selected wee given dosage of the second drug. The length of time in minutes for the drugs to reach a specified level in the blood was recorded for both drugs. The means and standard deviations of the two samples are recorded as follows:

| **Drug (1)** | **Drug (2)** |
|---|---|
| X = 10.1 | X = 8.9 |
| s = 4.2 | s = 3.8 |

Use a 5% level of significance to test the hypothesis that there is no difference in the mean time required for bodily absorption of these two drugs.

7. Comparing two population means

An industrial engineer consultant has conducted a time and motion study on a particular manufacturing assembly operation which he claims would save time. The production manager decides to test new procedure to see if it actually reduces the average assembly time. A random sample of ten assemblers is selected and each assembler is timed using the old procedure. Then the same assemblers are given training in the new procedures and are timed again as they perform the same operation. The following table shows the time in minutes taken for the operation under previous procedure and the new procedure:

| **Assembler** | **Old Procedure** | **New Procedure** |
|---|---|---|
| 1 | 10.6 | 10.0 |
| 2 | 6.7 | 7.4 |
| 3 | 8.0 | 6.1 |
| 4 | 9.5 | 6.0 |
| 5 | 7.1 | 7.1 |
| 6 | 7.0 | 6.0 |
| 7 | 6.4 | 5.5 |
| 8 | 6.9 | 7.0 |

| | | |
|---|---|---|
| 9 | 9.0 | 8.6 |
| 10 | 10.0 | 9.6 |

Assuming that the times under the old procedure, times under the new procedures and hence the data pf paired differences are all normally distributed, can we conclude at 99% confidence level that the new procedure reduces the average time required for the operation?