Introductory Applied Statistics

Dr. AbuBakr
A. AbdulMotaal

2023-2024

Faculty of Commerce
South Valley University

Introductory Applied Statistics

Dr. AbuBakr A. AbdulMotaal

Faculty of Commerce
South Valley University
Department of Quantitative Methods

2023 / 2024

Preface

In this book “Introduction to Applied Statistics”, mathematical background needed is basic arithmetic and basic elements of algebra. The primary purpose of the book is to take the mystery out of the subject matter and to present and explain this field of study in a manner which captures the student’s imagination in utilizing the statistical tools for the purpose of business decision making. The text has been written for facilitating usage by all business and economics majors.

Each topic in each chapter is explained by use of solved examples within the chapter so as to demonstrate the applicability of statistical tools described and learned in the chapter. There are additional unsolved problems at the end of each chapter. Unsolved problems are added at the end of each chapter so that the students acquire a reasonable degree of specialization in statistical thinking, decision making and problem solving.

The book covers various aspects of statistics in six chapters. All of these chapters deal with ‘Inferential Statistics’. In the first three chapters, various statistical terms and concepts are explained and analysis of Probability and its applications are discussed and explained in details. Chapters four, five and six enable the researchers to make decisions about populations using the results of samples taken from these populations.

AbuBakr A. AbdulMotaal

January, 2024

Contents

Chapter (1)
Introduction to Probability

Contents

Chapter 1
Introduction to Probability

In Statistics, probability plays a very important role where students should be able to have a good and clear idea about it. In this chapter, we will discuss probability as well as probability distribution. By having a look at the different probability formulas, it would be possible to get the perfect idea about it. We will also try to look forward to the different types of probability as well.

1.1 What is probability?

The meaning of probability is a possibility. It is a very interesting section of Statistics that deals with the occurrence of a random event. When it comes to the expression of the value, it is between zero and one. It helps in predicting as to how likely events are to happen. It is very important to get the perfect knowledge of the total number of outcomes in order to find the probability of a single event to occur.

Definition:

Theoretical Probability (Classical or A Priori Probability):

Moving forward to the theoretical probability which is also known as classical probability or priori probability we will first discuss about collecting all possible outcomes and equally likely outcome. When an experiment is done at random we can collect all possible outcomes.

Probability Formula:

When it comes to the probability formula, it is as under**:**

Probability of event to happen = P(E)

Where $$\mathbf{(E) = \frac{Number\ of\ Favorable\ Outcomes}{Total\ Number\ of\ Outcomes}}$$

It is to be kept in mind not to confuse the term “desirable outcome” with “favorable outcome” as both are different terms.

1.2 Events and Their Probabilities

Random Experiment:

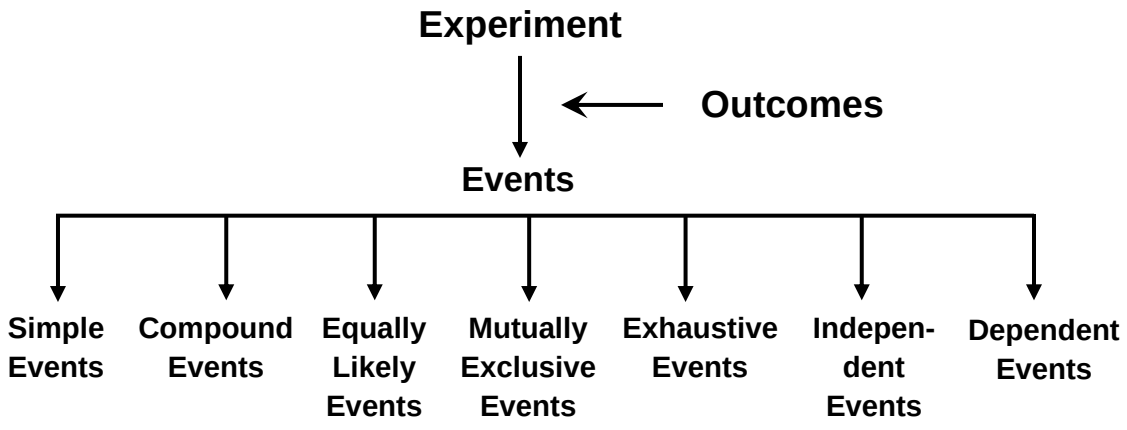
An experiment is any activity with an observable result. Tossing a coin, rolling a die or choosing a card from a pack of playing cards are all considered experiments an experiment is random if although it is repeated in the same manner every time, can result in different outcomes**:**

- The set of all possible outcomes is completely determined before carrying it out.
- Before we carry it out, we cannot predict its outcome.
- It can be repeated as many times as we want always under the same conditions (leading to different outcomes).

Outcomes of Experiment:

Outcome (or Sample Point) is the result of the experiment. The set of all possible outcomes or sample points of an experiment is called the Sample Space.

In a random experiment, the following types of events are possible**:**



Statistical Event:

Statistical Event is a subset of the sample space.

Playing Cards Probability:

Playing cards probability problems based on a well - shuffled deck of 52 cards.

Basic Concept on Drawing a Card:

- In a pack or deck of 52 playing cards, they are divided into **4** suits of 13 cards each, i.e. spades ♠ hearts ♥, diamonds ♦, clubs ♣.
- Cards of Spades and Clubs are black cards.
- Cards of Hearts and Diamonds are red cards.
- The card in each suit, are ace, king, queen, jack, 10, 9, 8, 7, 6, 5, 4, 3 and 2 and ace.
- King, Queen and Jack are face cards. So, there are 12 face cards in the deck of 52 playing cards.

Example (1.1):

A card is drawn from a well shuffled pack of 52 cards. Find the probability of**:**

(a) '2' of hearts **(b)** A king of red colour

(c) A black face card **(d)** A non-ace

(e) A non-face card of black colour

(f) Neither a queen nor a jack

Solution:

In a playing card there are 52 cards.

Therefore the total number of possible outcomes = 52

(a) '2' of hearts:

Number of favorable outcomes, i.e. '2' of hearts is 1 out of 52 cards.

Therefore, probability of getting '2' of heart is:

P('2' of heart) = 1/52

(b) A king of red colour:

Number of favorable outcomes i.e. 'a king of red colour' is 2 out of 52 cards.

Therefore, probability of getting a king of red colour is:
P(Red colour) = 2/52 = 1/26

(c) A black face card:

Cards of Spades and Clubs are black cards.
Number of face card in spades (king, queen and jack) = 3
Number of face card in clubs (king, queen and jack) = 3
Therefore, total number of black face card out of 52 cards
= 3 + 3 = 6
Therefore, probability of getting 'a black face card' is:
P(black face card) = 6/52 = 3/26

(d) A non - ace:

Number of ace cards in each of four suits namely spades, hearts, diamonds and clubs = 1
Therefore, total number of ace cards out of 52 cards = 4
Thus, total number of non-ace cards out of 52 cards
= 52 – 4 = 48
Therefore, probability of getting 'a non - ace' is:
P(None - ace card) = 48/52 = 12/13

(e) A non - face card of black colour:

Cards of spades and clubs are black cards.
Number of spades = 13
Number of clubs = 13
Therefore, total number of black cards out of 52 cards
= 13 + 13 = 26
Number of face cards in each suit namely spades and clubs
= 3 + 3 = 6
Therefore, total number of non - face card of black colour out of 52 cards = 26 - 6 = 20
Therefore, probability of getting 'non - face card of black colour' is equal to 20/52 = 5/13

(f) Neither a queen nor a jack:

Number of favorable outcomes for the event

= number of cards which are neither a queen nor a jack

= 52 - 8, [Since there are 4 queens and 4 jacks] = 44

Therefore, by definition,

P(Neither a queen or a jack) = 44/52 = 11/13

Example (1.2):

A card is drawn at random from a well - shuffled pack of cards numbered 1 to 20. Find the probability of

(a) Getting a number less than 7

(b) Getting a number divisible by 3.

Solution:

(a) Total number of possible outcomes = 20 (since there are cards numbered 1, 2, 3, ..., 20).

Number of favorable outcomes for the event E

= number of cards showing less than 7 = 6

(namely 1, 2, 3, 4, 5, 6).

P(Getting a number less than 7) = 6/20 = 3/10

(b) Total number of possible outcomes = 20.

Number of favorable outcomes for the event =

Number of cards showing a number divisible by 3 = 6 (namely 3, 6, 9, 12, 15, 18).

So, P(A number divisible by 3) = 6/20 = 0.3

Coin Toss Probability:

Coin Toss Probability are explained here with different examples. When we flip a coin there is always a probability to get a head or a tail is 50 percent.

Suppose a coin tossed, then we get two possible outcomes either a ‘head’ (**H**) or a ‘tail’ (**T**), and it is impossible to predict whether the result of a toss will be a ‘head’ or ‘tail’.

The probability for equally likely outcomes in an event is**:**

$$\mathbf{P(E) = \frac{Number\ of\ Favorable\ Outcomes}{Total\ Number\ of\ Possible\ Outcomes}}$$

Total number of possible outcomes = 2

(i) If the favorable outcome is head (H).
Number of favorable outcomes = 1.

Therefore, P(getting a head) = ½.

(ii) If the favorable outcome is tail (T).
Number of favorable outcomes = 1.

Therefore, P(getting a tail) = ½

Example (1.3):

A coin is tossed twice at random. What is the probability of obtaining:

(a) at least one head **(b)** the same face?

Solution:

The possible outcomes are: **HH , HT , TH , TT**.
So, total number of outcomes = 4.

(a) Number of favorable outcomes for this event
= Number of outcomes having at least one head
= 3 (as HH, HT, TH are having at least one head).

So, by definition, P(At least one head) = ¾

(b) Number of favorable outcomes for this event
= Number of outcomes having the same face
= 2 (as HH, TT are have the same face)

So, by definition, P(Same face) = 2/4 = ½

Example (1.4):

Three coins are tosses. Find the following for each of the probability two events (A and B), where

A = Getting exactly two heads

B = Getting at least two heads

Solution:

All possible outcomes **(8 outcomes)** are**:**

HHH , HHT , HTH , THH , HTT , THT , TTH , TTT

Event (A): Favorable outcomes for this event (3 outcomes) are**:**

HHT, HTH , THH

Therefore, P(A) = 3/8

Event (B): Favorable outcomes for this event (4 outcomes) are**:**

HHT, HTH ,THH , HHH

Therefore, P(B) = 4/8 = 0.5

Rolling A Die Probability:

Since the **die** is fair, each number in the set occurs only once. In other words, the frequency of each number is 1. To **determine the probability of rolling** any one of the numbers on the **die**, we divide the event frequency **(1)** by the number of all possible outcomes **(6)**, resulting in a **probability** of 1/6 for each of the 6 numbers (1 to 6). That is,

$$\mathbf{P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = \frac{1}{6}}$$

Example (1.5):

A 6 - sided die is thrown. What will be the probability of getting**:**

(a) A prime number?

(b) The number shown on the die is not a multiple of 2.

(c) The number shown on the die is greater than 4.

(d) The number shown on the die is 7.

Solution:

The total number of possible outcomes is 6.

The numbers of all possible outcomes are {1 , 2 , 3 , 4 , 5 , 6}

(a) A prime number:

The prime numbers among these are {2 , 3 , 5}

Probability of getting a prime number is**:**

$$\text{P(Prime number)} = \frac{\text{Favorable Outcomes}}{\text{Total Possible Outcomes}} = \frac{3}{6} = \frac{1}{2}$$

(b) Not a multiple of 2:

The numbers of all possible outcomes are {1, 2, 3, 4, 5, 6}

The numbers (not a multiple of 2) among these are {1,3, 5}

$$\text{P(Not a multiple of 2)} = \frac{\text{Favorable Outcomes}}{\text{Total Possible Outcomes}} = \frac{3}{6} = \frac{1}{2}$$

(c) The number is greater than 4:

The numbers of all possible outcomes are {1, 2, 3, 4, 5, 6}

The numbers (greater than 4) among these are {5, 6}

Probability of getting a number greater than 4 is**:**

$$\text{P(Even number)} = \frac{\text{Favorable Outcomes}}{\text{Total Possible Outcomes}} = \frac{2}{6} = \frac{1}{3}$$

(d) The number shown on the die is 7:

The numbers of all possible outcomes are {1, 2, 3, 4, 5, 6}

The numbers of 7 among these are {0}

Probability of getting a 7 is**:**

$$\text{P(Even number)} = \frac{\text{Favorable Outcomes}}{\text{Total Possible Outcomes}} = \frac{0}{6} = 0$$

Getting a 7 is an impossible Event

Rolling Two Dice Probability:

The possible outcomes of rolling two dice (A and B) are represented in the table below. Note that the number of total possible outcomes is equal to the number of all possible outcomes of the first die (6) multiplied by number of all possible outcomes of

the second die (6) , which is 36. We use notation like (1 , 2) to mean “1 on Die A, 2 on Die B”.

		Die (B)					
		1	**2**	**3**	**4**	**5**	**6**
Die (A)	**1**	**(1,1)**	**(1,2)**	**(1,3)**	**(1,4)**	**(1,5)**	**(1,6)**
	2	**(2,1)**	**(2,2)**	**(2,3)**	**(2,4)**	**(2,5)**	**(2,6)**
	3	**(3,1)**	**(3,2)**	**(3,3)**	**(3,4)**	**(3,5)**	**(3,6)**
	4	**(4,1)**	**(4,2)**	**(4,3)**	**(4,4)**	**(4,5)**	**(4,6)**
	5	**(5,1)**	**(5,2)**	**(5,3)**	**(5,4)**	**(5,5)**	**(5,6)**
	6	**(6,1)**	**(6,2)**	**(6,3)**	**(6,4)**	**(6,5)**	**(6,6)**

Example (1.6):

consider rolling two fair dice (A and B). Find the probability of**:**
Two dice are thrown simultaneously. Find the probability of:

(a) Getting six as a product. **(b)** Obtaining a doublet.

(c) Getting a sum of 8. **(d)** Getting a doublet of even numbers.

(e) Obtaining a multiple of 2 on one die and a multiple of 3 on the other die.

Solution:

For rolling two dice, the total number of possible outcomes is 6.

(a) Getting a six as a product:

Let E_1 = event of getting six as a product.

The numbers whose product is six will be

E_1 = [(1, 6), (2, 3), (3, 2), (6, 1)] = 4

Therefore, probability of getting ‘six as a product’ is

$$P(E_1) = \frac{\text{Number of Favorable Outcomes}}{\text{Total Number of Possible Outcomes}} = \frac{4}{36} = \frac{1}{9}$$

(b) Getting a doublet (equal numbers on both):

Let E_2 = event of getting a doublet.

The number which doublet will be:

E_2 = [(1 , 1) , (2 , 2) , (3 , 3) , (4 , 4), (5 , 5) , (6 , 6)] = 6

Therefore, probability of getting 'a doublet' is**:**

$$P(E_2) = \frac{\text{Number of Favorable Outcomes}}{\text{Total Number of Possible Outcomes}} = \frac{6}{36} = \frac{1}{6}$$

(c) Getting a sum of 8:

Let **E_3 = event of getting a sum of 8**.

The number which is a sum of 8 will be **E_3 = [(2 , 6) , (3 , 5) , (4 , 4), (5 , 3) , (6 , 2)]** = 5

So, probability of getting 'a sum of 8'**:**

$$P(E_3) = \frac{\text{Number of Favorable Outcomes}}{\text{Total Number of Possible Outcomes}} = \frac{5}{36}$$

(d) Getting a doublet of even numbers:

Let E_4 = event of getting a doublet of even numbers.

The events of a doublet of even numbers will be:

E_4 = [(2 , 2) , (4 , 4), (6 , 6)] = 3

Therefore, probability of getting 'a doublet of even numbers' will be

$$P(E_4) = \frac{\text{Number of Favorable Outcomes}}{\text{Total Number of Possible Outcomes}} = \frac{3}{36} = \frac{1}{12}$$

(e) Getting a multiple of 2 on one die and a multiple of 3 on the other die:

Let E_5 = event of getting a multiple of 2 on one die and a multiple of 3 on the other die.

Multiples of 2: 2, 4, 6

Multiples of 3: 3, 6

The events of a multiple of 2 on one die and a multiple of 3 on the other die will be **E_5 = [(2 , 3) , (4 , 3) , (6 , 3) , (2 , 6) , (4 , 6) , (6 , 6) , (3 , 2) , (3 , 4), (6 , 3) , (6 , 2) , (6 , 4) = 11**

So, the probability of getting ‘a multiple of 2 on one die and a multiple of 3 on the other die’ will be:

$$P(E_5) = \frac{\text{Number of Favorable Outcomes}}{\text{Total Number of Possible Outcomes}} = \frac{11}{36}$$

1.3 Some Basic Relationships of Probability

Impossible Event:

An event that cannot happen.

Certain Event:

Is the event that will always happen. It is also called **‘sure event’**.

Exhaustive Events:

In probability theory and logic, a set of events is jointly or collectively exhaustive if at least one of the events must occur. For example, when rolling a six-sided die, the events 1, 2, 3, 4, 5, and 6 are collectively exhaustive, because they include the entire range of possible outcomes.

Complementary of an Event:

The event ‘not E' is called complementary event of the event E. If E occurs, its compliment is $\bar{E}$ which does not occur.
Compliment of an event is denoted by $\bar{E}$ or E^c. This means that**:**

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

Therefore: $P(E) = 1 - P(\bar{E})$

Or $P(\bar{E}) = 1 - P(E)$

For example:

1. When a coin is tossed, getting ‘head’ and getting ‘tail’ are complimentary events of each other.
2. When two coins are tossed, getting ‘at least one head’ and getting ‘no head’ are complimentary event of each other.

3. For selecting a card from a pack of playing cards**: '**Getting a nine' and 'not getting a nine' are complimentary event of each other.

4. For a student**:**

- 'passing the exam' and 'failing the exam' are complimentary events of each other.
- 'Answering a given question correctly' and 'answering the same question incorrectly' are complimentary events of each other.

Example (1.7):

A bag contains red and white balls. The probability of getting a red ball from the bag of balls is 1/6. What is the probability of not getting a red ball?

Solution:

The probability of getting a red ball from the bag of balls is 1/6.
Therefore, the probability of not getting a red ball is**:**
P(ball is not red) = 1 – 1/6 = 5/6
Therefore, the probability of not getting a red ball is 5/6.

Simple Event:

A simple event is one that can only happen in one way - in other words, it has a single outcome.
Examples for Simple Event (with only one outcome)**:**
• The probability of rolling a 3 on a die. (1/6)
• The probability of tossing a head with a coin (1/2).

Compound Event:

A compound event is more complex than a simple event, as it involves the probability of more than one outcome.

For example:

- The probability of rolling an even number less than 5 on a die (2/6).
  [Even numbers {2 , 4 , 6}; less than 5 {1 , 2 , 3 , 4}; both {2 , 4}]

- The probability of drawing a red 8 from a deck of cards (2/52). [red card {26 cards}; eight {4 cards}; both {red 8 of hearts, red 8 of diamonds}].
- The probability of rolling an even number on a die, then tossing a head on a coin. (3/12). Even number (2 , 4 , 6);
- The probability of tossing three coins and getting at least 2 heads (4/8).

Mutually Exclusive Events:

If two events are such that they cannot occur simultaneously for any random experiment are said to be mutually exclusive events.

For two mutually exclusive events (A and B):

$$P(A \text{ and } B) = 0$$

Where P(A and B) is the probability that the two events A and B occurring together (simultaneously).

Which means that the occurrence of one of them excludes the occurrence of the other event.

For example:

- Events in tossing of a die are “even face” and “odd face” which are known as mutually exclusive events.
- In rolling a coin**:** The two events ‘obtaining a Head’ and ‘obtaining a Tail’ are mutually exclusive events.
- In selecting a card from a pack of playing cards**: ‘**Getting a jack’ and ‘getting a seven’ are mutually exclusive events.

Mutually Non - Exclusive Events:

Two events A and B are said to be mutually non - exclusive events if both the events A and B have at least one common outcome between them. They are also known as **‘compatible events’**.

The events A and B cannot prevent the occurrence of one another, so from here we can say that the events A and B have something common in them. This gives:

$$P(A \text{ and } B) \neq 0$$

For example:

- In the case of rolling a die the event of getting an 'odd-face' and the event of getting 'less than 4' are not mutually exclusive events.
  The event of getting an 'odd-face' and the event of getting 'less than 4' occur when we get either 1 or 3.
  Let 'X' is denoted as event of getting an 'odd-face' and 'Y' is denoted as event of getting 'less than 4'
  The events of getting an odd number (X) = {**1**, **3**, 5}
  The events of getting less than 4 (Y) = {**1**, 2, **3**}
  But "odd-face" and "multiple of 3" are not mutually exclusive, because when "face-3" occurs both the events "odd face" and "multiple of 3" are said to be occurred simultaneously.
  We see that two simple-events are always mutually exclusive while two compound events may or may not mutually exclusive.
- In selecting a card from a pack of playing cards**: '**Getting a Jack' and 'getting a card of red color' are not mutually exclusive events. This is because there is a Jack of red color.

Addition Law of Probabilities:

- **For Non-mutually Exclusive Events:**

  The addition law of probability (sometimes referred to as the addition rule or sum rule), states that the probability that A or B will occur is the sum of the probabilities that A will happen and that B will happen, minus the probability that both A and B will happen. The addition rule is summarized by the formula**:**

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

Example (1.8):

A card is drawn from a deck of 52 playing cards, what is the probability of getting a heart or a face card (king, queen, or jack)?

Solution:

Let H denotes drawing a heart and F denotes drawing a face card. Since there are 13 hearts and a total of 12 face cards (3 of each suit**:** spades, hearts, diamonds and clubs), but only 3 face cards of hearts, we obtain**:**

P(H) =13/52, P(F) = 12/52, P(F and H) = 3/52

Using the addition rule, we get**:**

$$P(H \text{ or } F) = P(H) + P(F) - P(H \text{ and } F)$$
$$= 13/52 + 12/52 - 3/52$$
$$= 22/52 = 11/26$$

- **For Mutually Exclusive Events:**
  Suppose A and B are mutually exclusive events, Then the probability of A and B is Zero. So,

  P(A and B) = 0.

The addition law then simplifies to**:**

P(A or B) = P(A) + P(B)

Example (1.9):

Two dice (A and B0 are thrown. Find the following probability:
P(A = 3 or A + B = 8).

Solution:

The events of getting (A = 3) are: **(3,1) , (3,2) , (3,3) , (3,4) , (3,5) , (3,6)**

The events of getting (A + B = 8) are: **(2,6) , (3,5) , (4,4) , (5,3) , (6,2)**

Both: A = 3 and A + B = 8 is: **(3,5)**

Thus, applying the formula of the addition rule we get:

$$P(A = 3 \text{ or } A + B = 8)$$
$$= P((A = 3) + P(A + B = 8) - P(A = 3 \text{ and } A + B = 8)$$
$$= 6/36 + 5/36 – 1/36 = 10/36$$
$$= 5/18$$

Example (1.10):

Suppose a card is drawn from a deck of 52 playing cards: what is the probability of getting a five or a ten?

Solution:

Let F represent the event that a five is drawn and T represent the event that a ten is drawn. These two events are mutually exclusive, since there are no fives that are also tens. Thus**:**

$$P(F \text{ or } T) = P(F) + P(T) = 4/52 + 4/52$$
$$= 8/52 = 2/13$$

Multiplication Law of Probability:

Dependent Events:

When two events are dependent events, one event influences the probability of another event. A dependent event is an event that **relies on another event** to happen first. More formally, we say that when two events are dependent, the occurrence of one event influences the probability of another event.

Independent Events:

In statistics and probability theory, independent events are two events wherein the occurrence of one event does not affect the occurrence of another event or events. The simplest example of such events is tossing two coins. The outcome of tossing the first coin cannot influence the outcome of tossing the second coin. When two events are independent, one event does not influence the probability of another event.

Dependent or Independent?

Card example:

Cards are often used in probability as a tool to explain how one seemingly independent event can influence another. For example, if you choose a card from a deck of 52 cards, your probability of getting a Jack is 4 out of 52. Statistically, you can write it like this**:**

$$\text{P(Jack)} = \frac{\text{Number of Jacks in a deck of cards}}{\text{Total number of cards in a deck}} = \frac{4}{52} = \frac{1}{13}$$

If you replace the jack and choose again (assuming the cards are shuffled), the events are independent. Your probability remains the same (1/13). Choosing a card over and over again would be an independent event, because each time you choose a card (a "trial" in probability) it's a separate, independent event.

But what if the card was **kept out** of the pack? Let's say you pulled the three of hearts, but you're still searching for that jack. The second time you pull out a card, the deck is now 51 cards, so**:**

$$\text{P(Jack)} = \frac{\text{Number of Jacks in a deck of cards}}{\text{Total number of cards in a deck}} = \frac{4}{51}$$

The probability has increased from 4/52 (with replacement of the jack) to 4/51 (the jack isn't replaced), so choosing cards in this manner is an example of a **dependent event**.

Independent events are frequently confused with mutually exclusive events. However, they are two distinct concepts. Mutually exclusive events are events that cannot occur simultaneously. The concept of independent events is not related to the simultaneous occurrence of the events, but it is only concerned with the influence of the occurrence of one event on another.

Multiplication Law for Independent Events:

The law of multiplication is used when we want to find the probability of events occurring (it is also known as the joint probability of independent events). The rule of multiplication states the following**:**

$$\mathbf{P(A \text{ and } B) = P(A) \times P(B)}$$

Where P(A and B) is the probability that the two events A and B occurring together (simultaneously).

In other words, if you want to find the probability of both events A and B taking place, you should multiply the individual probabilities of the two events.

For example, if we toss a coin twice, then the probability that the first toss results in a head and the second toss results in a tail is given by**:**

$$P(H \text{ and } T) = P(H) \times P(T)$$
$$= \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

Multiplication Rule for Dependent Events:

If Events A and B are not independent, meaning that the probability of occurrence of an event is dependent or conditional upon the occurrence or non-occurrence of the other event, then the probability that they will both occur is given by**:**

$$\mathbf{P(A \text{ and } B) = P(A) \times P(B \mid \text{given the occurrence of A})}$$

This formula is written as:

$$\mathbf{P(A \text{ and } B) = P(A) \times P(B \mid A)}$$

Or, altering the notation,

$$\mathbf{P(A \text{ and } B) = P(B) \times P(A \mid B)}$$

Where **P(B | A)** means the probability of event B on the condition that event A has occurred, and

P(A | B) means the probability of event A on the condition that event B has occurred.

Example (1.11):

A box has 6 black balls and 4 white balls. A ball is drawn at random from the box. Then a second ball is drawn without replacement of the first ball back in the box. What is the probability that both these balls are black?

Solution:

The probability of the second ball being black or white would depend upon the result of the first draw as to whether the first ball was black or white.

The probability that both balls are black is given by**:**

P(Two black balls)

= P(black on 1^{st} draw) × P(black on 2^{nd} draw | black on 1^{st} draw)

= 6/10 × 5/9 = 1/3

This is because, first there are 6 black balls out of total 10, but if the first ball drawn is black then we are left with 5 black balls out of a total of 9 balls.

Conditional Probability:

In many situations, a manager may know the outcome of an event that has already occurred and may want to know the chances of a second event occurring based upon the knowledge of the outcome of the earlier event. We are interested in finding out as to how additional information obtained as a result of the knowledge about the outcome of an event affects the probability of the occurrence of the second event.

The conditional probability of an event **A** is the probability that the event will occur given the knowledge that an event B has already occurred. This probability is written P(A | B), notation for the probability of A given B.

If A and B are two dependent events, then the conditional probability of A given B is defined as**:**

$$\mathbf{P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)}, \text{ where } P(B) > 0}$$

Or, $$\mathbf{P(B \mid A) = \frac{P(A \text{ and } B)}{P(A)}, \text{ where } P(A) > 0}$$

Where **P(A | B)** is interpreted as the probability of event A on the condition that event B has occurred, and

P(B | A) is interpreted as the probability of event B on the condition that event A has occurred.

Example (1.12):

Two dice (A and B) are thrown. Find the probability of**:**

(a) Getting a sum of 8, given that the first die lands on 3. That is,

$$P((A + B) = 8 \mid A = 3)$$

(b) Getting a 3 on Die A, given a sum of 8. That is,

$$P(A = 3 \mid (A + B) = 8)$$

Solution:

(a) Let **S8** be the event that the sum is **8**, and
A3 is the event of **3** on Die **A**.

The events of a 3 on Die A will be**:**

A3 = [(3,1), (3,2), (3,3), (3,4), (3,5), (3,6)] = 6

Thus, P(A3) = 6/36 = 1/6

The event of a sum of 8 is: **(2,6) , (3,5) , (4,4) , (5,3) , (6,2)**

Thus, P(S8) = 5/36

The events of a sum of 8 and a 3 on Die A will be**:**

A3 = [(3, 5)] = 1

Thus, P(S8 and A3) = 1/36

The required probability is**: P(S8 | A3)**

Applying the formula of the conditional probability, we get**:**

$$P(S8 \mid A3) = \frac{P(S8 \text{ and } A3)}{P(A3)} = \frac{1/36}{1/6} = \frac{1}{6}$$

(b) Here, the required probability is: **P[(A = 3) | (A + B) = 8]**

The events of a sum of 8 are already obtained.

$$\text{Then, } P(A3 \mid S8) = \frac{P(S8 \text{ and } A3)}{P(S8)} = \frac{1/36}{5/36} = \frac{1}{5}$$

In the case where events A and B are independent (where event A has no effect on the probability of event B), the conditional probability of event A given event B is simply the probability of event A, that is P(A). In other words, If two events are independent, the probabilities of their outcomes are not dependent on each other. Therefore, the conditional probability of two independent events A and B is**:**

$$\mathbf{P(A|B) = P(A)}$$

Or

$$\mathbf{P(B|A) = P(B)}$$

The probability of A, given that B has happened, is the same as the probability of A. Likewise, the probability of B, given that A has happened, is the same as the probability of B. This shouldn't be a surprise, as one event doesn't affect the other.

You can use the following equation to figure out probability for independent events:

$$\mathbf{P(A \text{ and } B) = P(A) \times P(B)}$$

1.4 Bayes' Theorem:

In this section we extend the discussion of conditional probability to include applications of Bayes' theorem (or Bayes' rule), which we use for revising a probability value based on additional information that is later obtained. One key to understanding the essence of Bayes' theorem is to recognize that we are dealing with sequential events, whereby new additional information is

obtained for a subsequent event, and that new information is used to revise the probability of the initial event. In this context, the terms prior probability and posterior probability are commonly used.

Definitions:

Prior Probability:

A prior probability is an initial probability value originally obtained before any additional information is obtained.

Posterior Probability:

A posterior probability is a probability value that has been revised by using additional information that is later obtained.

The theorem is named after English statistician, **Thomas Bayes**, who discovered the formula in 1763. It is considered the foundation of the special statistical inference approach called the Bayes' inference'.

Besides Statistics, the Bayes' theorem is also used in various disciplines, with medicine and pharmacology as the most notable examples. In addition, the theorem is commonly employed in different fields of finance.

The Bayes' theorem is expressed in the following formula**:**

$$\mathbf{P(A|B)} = \frac{\mathbf{P(A)P(B|A)}}{\mathbf{P(A)P(B|A) + P(\bar{A})P(B|\bar{A})}}$$

Where P(A | B) is the probability of event A, given B has occurred.
and P(B | A) is the probability of event B, given A has occurred.
P(A) is the probability of event A.
$P(\bar{A}) = 1 - P(A)$,
P(B) is the probability of event B.

Split Event Rule:

If**:** $\mathbf{P(B) = P(A_1)P(B|A_1) + P(A_2)P(B|A_2)}$

Where
The events $\mathbf{A_1}$ and $\mathbf{A_2}$ are **mutually exclusive** events,
The event **B** is called **'Split Event'**, and
The **formula** named as **'Split Event Rule'**.

Generalize Split Event Rule:

Let $A_1,\ldots\ldots,A_n$ be mutually exclusive (disjoint) events whose union is the whole of the sample space (partition) and assume $P(A_i) > 0$ for every i. For every event B we have**:**

$$\mathbf{P(B) = \sum_{i=1}^{i=n} P(A_i)P(B \mid A_i)}$$

$$\mathbf{= P(A_1 \text{ and } B) + P(A_2 \text{ and } B) + \cdots\cdots + P(A_n \text{ and } B)}$$

$$\mathbf{= P(A_1)P(B|A_1)\ P(A_2)(PB|A_2) + \cdots\cdots + P(A_n)P(B|A_n)}$$

Where
$\mathbf{A_1, A_2, \ldots\ldots, A_n}$ are **mutually exclusive** events,
B is a **'Split event'**,
n = number of mutually exclusive events
This formula named as **'Generalized Split Event Rule'**.

Example (1.13):

Bag (1) contains 4 white and 6 green balls while **Bag (2)** contains 3 white and 2 green balls. One ball is drawn at random from one of the bags, and it is found to be **green**.

(a) Find the probability that it was drawn from **Bag (1)**.
(b) What is the probability that it was drawn from **Bag (2)**?

Solution:

(a) Let **B1** be the event of choosing the **Bag (1)**,
B2 the event of choosing the **Bag (2)**, and
G be the event of drawing a green ball.

Then, P(B1) = P(B2) = 1/2

Also, P(G|B1) = P(Drawing a green ball from Bag (1) = 6/10 = 3/5,

P(G|B2) = P(Drawing a green ball from Bag (2) = 2/5

By using Bayes' theorem, the probability of drawing from Bag (1), given it was green is as follows**:**

The required probability is**:** P(B1|G)

$$P(B1|G) = \frac{P(B1)P(G|B1)}{P(G)}$$

P(G) = P(B1)P(G|B1) + P(B2)P(G|B2)

=1/2 × 3/5 + 1/2 × 2/5 = 5/10 = 0.5

$$P(B1|G) = \frac{P(B1)P(G|B1)}{P(G)} = \frac{(\frac{1}{2})\times(\frac{3}{5})}{\frac{1}{2}} = 0.6$$

(b) P(B1|G) = 1 - P(B2|G) = 1 – 0.6 = 0.4

Example (1.14):

A man is known to speak truth 2 out of 3 times. He throws a die and reports that the number obtained is a four. Find the probability that the number obtained is actually a four.

Solution:

Let A be the event that the man reports that number four is obtained.

Let E1 be the event that four is obtained and.

E2 be its complementary event.

Then, P(E1) = Probability that four occurs = 1/6

P(E2) = Probability that four does not occurs

= 1 – P(E1) = 1 −1/6 = 5/6

Also, P(A | E1) = Probability that man reports four and it is actually a four = 2/3

P(A | E2) = Probability that man reports four and it is not a four = 1/3

By using Bayes' theorem, probability that number obtained is actually a four is:

$$P(E1 \mid A) = \frac{P(E1 \text{ and } A)}{P(A)} = \frac{P(E1)P(A \mid E1)}{P(A)}$$

$$P(A) = P(E1)P(A \mid E1) + P(E2)P(A \mid E2)$$

$$= 1/6 \times 2/3 + 5/6 \times 1/3 = 7/18$$

$$P(E1 \mid A) = \frac{\frac{1}{6} \times \frac{2}{3}}{\frac{1}{6} \times \frac{2}{3} + \frac{5}{6} \times \frac{1}{3}} = \frac{2}{7}$$

1.5 Tree Diagram:

A tree diagram is simply a way of representing a sequence of events. Tree diagrams are particularly useful in probability since they record all possible outcomes in a clear and uncomplicated manner. The following two examples illustrates how Tree Diagram simplify calculations for probabilities.

Example (1.15):

Let's take a look at a simple example, flipping a coin and then rolling a die. We might want to know the probability of getting a Head and a 4.

Solution:

The following is a list of all the possible outcomes**:**

(H,1) , (H,2) , (H,3) , (H,4) , (H,5) , (H,6)

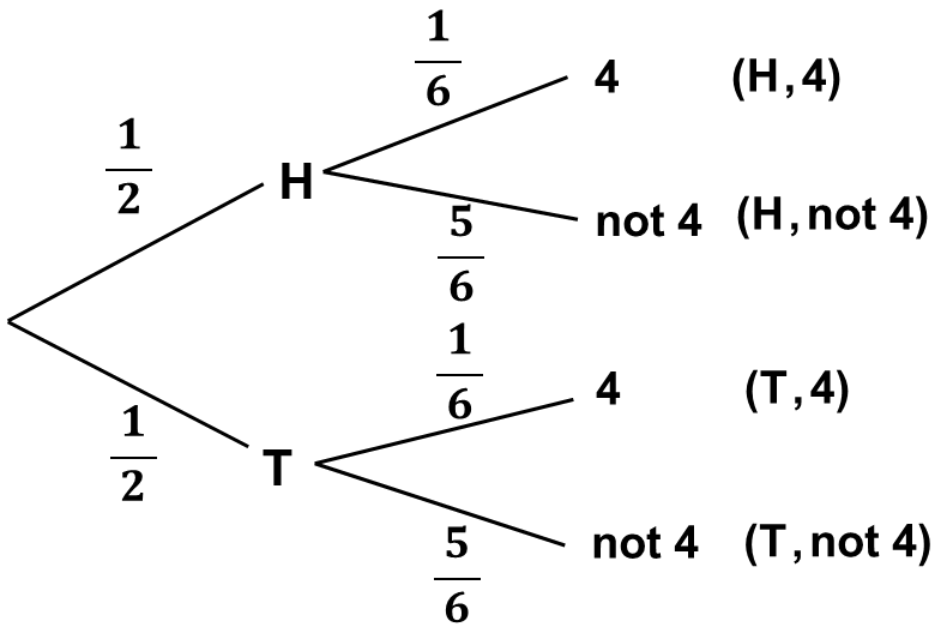
(T,1) , (T,2) , (T,3) , (T,4) , (T,5) , (T,6)

Probability of getting a Head and a 4:

P(H and 4) = (1/2)(1/6) = 1/12 as the two events are independent

Here is one way of representing the situation using a tree diagram. To save time, we have chosen not to list every possible

die throw (1 , 2 , 3 , 4 , 5 , 6) separately, so we have just listed the outcomes "4" and "not 4":



Each path represents a possible outcome, and the fractions indicate the probability of travelling along that branch. For each pair of branches the sum of the probabilities adds to 1.

"And" Means Multiply

So how might we work out P(H,4) from the tree diagram? We could word this as the probability of getting a Head and then a 4. This is the first path.

Half the time, I expect to travel along the first branch.

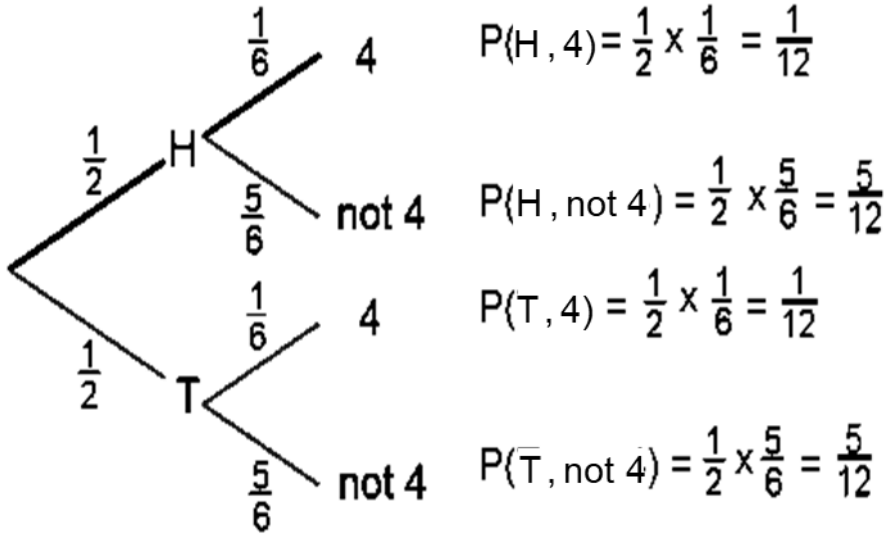
H	T

Then, on one sixth of those occasions, we will also travel along the second branch. We can think of this as 1/6 of 1/2.

$$\frac{1}{6} \text{ of } \frac{1}{2} = \frac{1}{6} \times \frac{1}{2} = \frac{1}{12}$$

1	1
2	2
3	3
4	4
5	5
6	6

So, this is why we said that you multiply across the branches of the tree diagram.



"And" only means multiply if events are independent, that is, the outcome of one event does not affect the outcome of another. This is certainly true for our example, since flipping the coin has no impact on the outcome of the die throw.

"Or" Means Add:

Example (1.16):

Now let's consider the probability of getting a Head or a 4 for the previous example.

Solution:

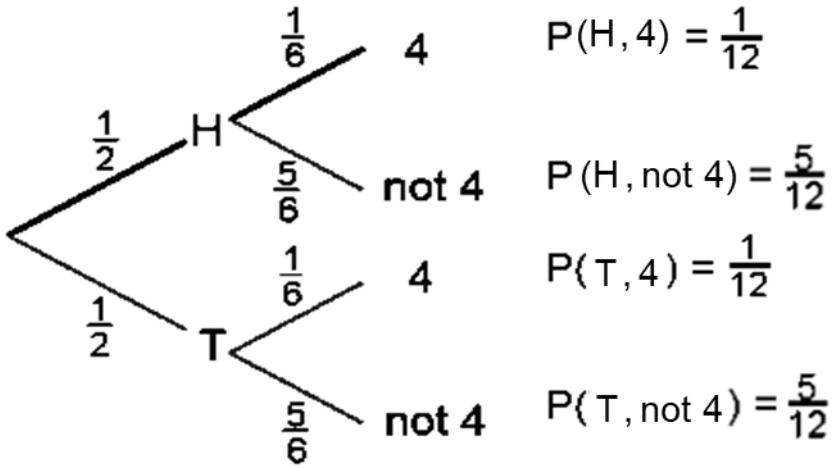
We are using the word "or" in its statistical sense to mean "Head or 4 or both", as opposed to the common usage which often means "either a Head or a 4"**:**

(H,1) , (H,2) , (H,3) , (H,4) , (H,5) , (H,6) , (T,4)

So P(H or 4) = P(H) + P(4) – P(H and 4)

= 1/2 + 1/6 – (1/2)(1/6) = 7/12

Again, we can work this out from the tree diagram, by selecting every branch which includes a Head or a 4:



Each of the ticked branches shows a way of achieving the desired outcome. So P(H or 4) is the sum of these probabilities:

P(H or 4) = P(H,4) + P(H, not 4) + P(T,4)

=1/12 + 5/12 + 1/12 = 7/12

So, this is why we said that you add down the ends of the branches.

Picturing the Probabilities:

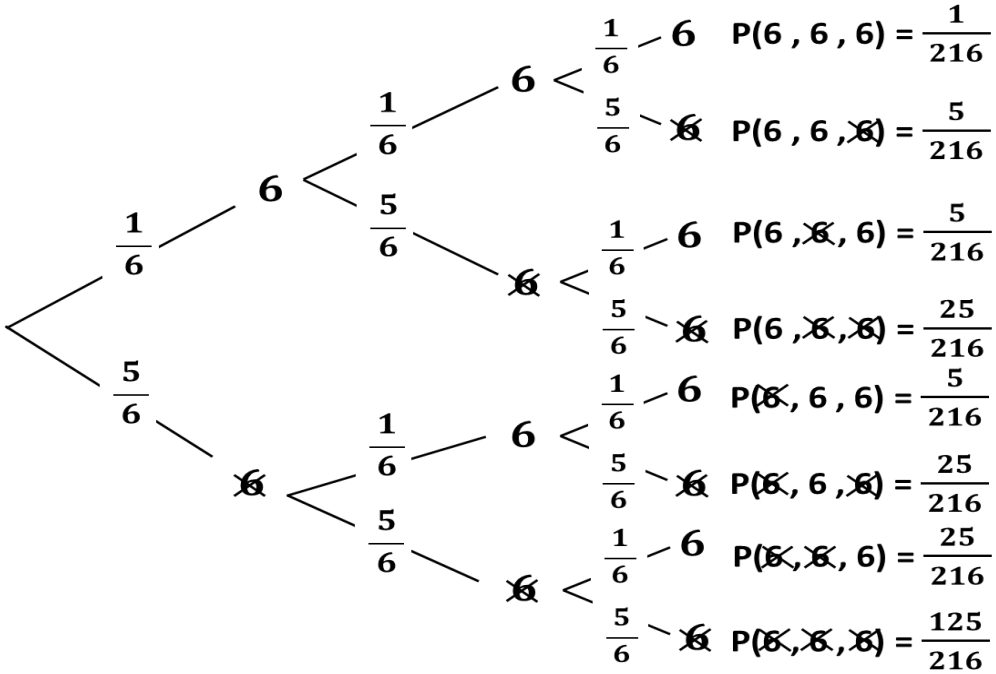
Example (1.17):

A die is thrown three times, find the probability of getting one, two or three sixes.

Solution:

We draw a tree diagram as shown below**:**

Check that you agree with the probabilities at the end of each branch before reading on.



We can now work out**:**

P(three sixes) = 1/216

P(exactly two sixes) = 5/216 + 5/216 + 5/216 = 15/216

P(exactly one six) = 25/216 + 25/216 + 25/216 = 75/216

P(no sixes) = 125/21/6

Again, check that you understand where these probabilities have come from before reading on.

To really check your understanding, think about the outcomes that contribute to each of the probabilities on the tree diagram. For example, P(6, not 6, 6) is 5/216, because out of the 216 total outcomes there are five outcomes which satisfy (6, not 6, 6) which are**:**

(6 , 1 , 6) , (6 , 2 , 6) , (6 , 3 , 6) , (6 , 4 , 6) , (6 , 5 , 6)

Now, you can know why there are 25 outcomes that satisfy (not 6, not 6, 6)?

What about the other probabilities on the tree diagram?
We hope these examples help you to understand what's happening next time you come across a tree diagram, and that it helps you to construct your own tree diagrams to solve problems.

Now, we will provide some other examples of general probabilities covering the concepts, rules, and theories that were previously introduced.

Example (1.18):

Suppose the probability that student (A) will pass the statistics exam is 0.8. This probability for student (B) is 0.7. Find the probability of:

(a) Both students will pass the exam?
(b) Student A fails the exam and student B passes the exam.
(c) Only one of them will pass the exam.
(d) At least one of them will fail the exam.

Solution:

(a) Both students will pass the exam:

Let A be the event that student A will pass the exam and B is the event that student B will pass the exam. That is,

$P(A) = 0.8$ and $P(B) = 0.7$

These two events are independent, since the occurrence of one of them does not affect the occurrence of the another event.

Therefore,

$$P(\text{Both will pass the exam}) = P(A \text{ and } B) = P(A) \times P(B)$$
$$= 0.8 \times 0.7$$
$$= 0.56$$

(b) Student A fails the exam and student B passes the exam:

$P(A \text{ fails}) = P(\bar{A}) = 1 - P(A) = 1 - 0.8 = 0.2$

$P(\bar{A} \text{ and } B) = P(\bar{A})P(B) = 0.2 \times 0.7 = 0.14$

(c) Only one of them will pass the exam:

The events of one of them will pass the exam are**: $A\bar{B}$, $\bar{A}B$**

$P(\bar{B}) = 1 - P(B) = 1 - 0.7 = 0.3$

$P(\text{one of them will pass the exam}) = P(A\bar{B}) + P(\bar{A}B)$

Since A and $\bar{B}$ are independent, the same for $\bar{A}$ and B,

$$P(A\bar{B}) + P(\bar{A}B) = P(A)P(\bar{B}) + P(\bar{A})P(B)$$
$$= 0.8 \times 0.3 + 0.2 \times 0.7 = 0.38$$

(d) At least one of them will fail the exam:

The events of at least one fails the exam are**: $A\bar{B}$, $\bar{A}B$, $\bar{A}\bar{B}$**

Note: $\bar{A}$ and $\bar{B}$ are also independent

Thus,

P(At least one of them will fail the exam)

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

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

$= 0.8 \times 0.3 + 0.2 \times 0.7 + 0.2 \times 0.3 = 0.44$

OR: Using the **complementary event**, we get**:**

P(At least one of them will fail the exam)

$= 1 - P(\text{Both pass the exam})$

$= 1 - P(AB) = 1 - P(A)P(B)$

$= 1 - 0.8 \times 0.7 = 0.44$

The same result obtained before.

Example (1.19):

Let X and Y are two independent events such that P(X) = 0.3 and P(Y) = 0.7. Find**:**

(a) P(X and Y) **(b)** P(X or Y) **(c)** P(X and $\bar{Y}$)

Solution:

Given P(X) = 0.3 and P(Y) = 0.7 and events X and Y are independent of each other.

(a) P(X and Y) = P(X) P(Y) = 0.3 × 0.7 = 0.21

(b) P(X or Y) = P(X) + P(Y) – P(X and Y)
= P(X) + P(Y) – P(X)P(Y)
= 0.3 + 0.7 – 0.21 = 0.79

(c) P(X and $\overline{Y}$) = P(X)P($\overline{Y}$) = 0.3 × 0.3 = 0.09

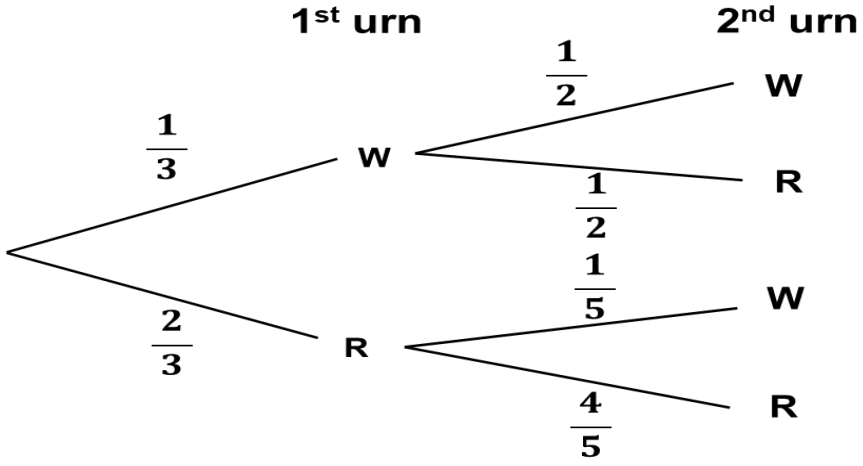
Example (1.20):

An urn contains 1 white ball and 2 red balls. One ball is taken from the urn at random. If it is a white ball, it is returned into the urn together with another white ball. If it is a red ball, it is returned together with other 2 red balls. In a second stage, another ball is taken from the urn at random.

(a) What is the probability that the second ball is white?

(b) What is the probability that the second ball is red?

Solution:



Events:

Let W1 ≡ 'first ball is white'; R1 ≡ 'first ball is red'
W2 ≡ 'second ball is white' R2 ≡ 'second ball is red'

Probabilities:

- P(W1) = 1/3; P(R1) = 2/3
- P(W2 | W1) = 1/2; P(R2 | W1)=1/2
- P(W2 | R1) = 1/5; P(R2 | R1) = 4/5

(a) P(Second ball is white) = P(W2)

= P(W1)P(W2 |W1) + P(R1)P(W2 |R1)

= (1/3)(2/4) + (2/3)(1/5)

= 1/6 + 2/15 = 0.3

(b) P(Second ball is white) = P(R2)

= P(W1)P(R2 |W1) + P(R1)P(R2 |R1)

= (1/3)(1/2) + (2/3)(4/5)

= 1/6 + 8/15 = 0.7

Or P(R2) = 1 – P(W2) = 1 – 0.3 = 0.7

Note that:

P(W2|R1) + P(R2|R1) = 1 Complementary probability

P(W2|W1) + P(R2|W1) = 1 Complementary probability

Example (1.21):

Refer to Example (1.19), If the second ball taken from the urn is red, what is the probability that the first one was also red?

Solution:

$$P(R1|R2) = \frac{P(R1 \text{ and } R2)}{P(R2)} = \frac{P(R1)P(R2|R1)}{P(R2)} = \frac{(\frac{2}{3})(\frac{4}{5})}{0.7}$$

= 0.762.

Exercises for Chapter 1
(Exam Questions)

(1.1) At a large factory, the employees were surveyed and classified according to their level of education and whether or not they smoked. The data are shown in the table.

Smoking Habit	Educational Level of Graduate		
	Preparatory School	Secondary School	College
Smoke	2	8	15
Do not smoke	8	32	35

If an employee is selected at random, find these probabilities**:**

(a) The employee did not graduate from college.

(b) The employee smokes and graduated from the secondary school.

(c) Given that the employee is a nonsmoker, he or she graduated from college.

(d) The employee smokes given he or she graduated from the preparatory school.

(e) The employee smokes or graduated from the secondary school.

(Exam 1998)

(1.2) A bag contains two black balls and two white balls. A ball is drawn and replaced by a ball of the opposite color. Then another ball is drawn from the bag. Find the probability that the first ball drawn was white, given that the second ball drawn was black.

(Exam 1999)

(1.3) Two teams **A** and **B** play a football match against each other. The probabilities for each team of scoring **0**, **1**, **2** goals are shown in the following table:

Number of Goals	0		1		2	
Team	A	B	A	B	A	B
Probability of Scoring	0.3	0.2	0.5	0.7	0.2	0.1

Calculate the probability of**:**

(a) A winning **(b)** a draw **(c)** B winning

(Exam 2000)

(1.4) Two persons shoot a target in the same time. The probability that the first hits the target is 0.8, the probability that the second hits the target is 0.9. Find the following probabilities**:**

(a) Hitting the target **(b)** The first person only hits the target

(c) At least one of them hits the target

(Exam 2001)

(1.5) Three identical bags each contain two balls. The balls are identical apart from their colour. One bag contains two white balls, the second two black balls and the third one black and one white ball. A bag is selected at random and a ball removed from it. If the ball is white, what is the probability that the other ball in the bag is also white?

(1.6) The following table gives the two-way classification of **100** randomly selected employees from a city based on gender and commuting time from home to work.

	Commuting Time		
Gender	**Less Than 30 Minutes**	**30 Minutes to One Hour**	**More Than One Hour**
Men	25	20	5
Women	18	20	12

(Exam 2002)↓

(a) If one employee is selected at random from these **100** employees, find the following probabilities**:**

(i) P(commutes for more than **one** hour or man).

(ii) P(woman given that she commutes for less than **30** minutes).

(iii) P(man or woman).

(b) Are the events "man" and "**30** minutes to one hour" independent**?** Why or why not**?**

(c) Are the events "woman" and "more than one hour" mutually exclusive**?** Why or why not**?**

(1.7) In a town it never rains on Friday, Saturday, Sunday or Monday. The probability that it rains on a given Tuesday is 1/5. For each of the remaining two days, Wednesday and Thursday, the conditional probability that it rains, given that it rained the previous day, is **α**, and the conditional probability that it rains, given that it did not rain the previous day, is **β**.

(a) Show that the (unconditional) probability of rain in a given Wednesday is $\frac{\alpha + 4\beta}{5}$, and find the probability of rain on a given Thursday.

(b) Explain the implications of the case **α = β**.

(1.8) A **die** is loaded so that the chance of throwing a **1** is $\frac{1}{4}$ **x**, the chance of a **2** is $\frac{1}{4}$, and the chance of a **6** is $\frac{1}{4}$ **(1 – x)**. The chance of a **3**, **4**, or **5** is $\frac{1}{6}$. The die is thrown twice.

(a) Prove that the chance of getting a total of **7** is $\frac{9x - 9x^2 + 10}{72}$.

(b) Find the value of **x** which will make this chance a maximum, and find this maximum probability.

(Exam 2002)

(1.9) A bag contains **three** red, **four** white, and **five** black balls. If **two** balls are taken, what is the probability that they are of the same color if**:**

(a) the first ball after being identified is put back in the bag before the second ball is selected.

(b) the first ball after being identified is removed before the second ball is selected.

(Exam 2003 Sohag)

(1.10) The probability that a January night will be icy is **0.25**. On an icy night, the probability that there will be a car accident in a certain dangerous corner is **0.04**. If it is not icy, the probability of an accident is **0.01**. What is the probability that:

(a) 12 January will be icy and there will be an accident.

(b) there will be an accident on 12 January.

(Exam 2003 Qena)

(1.11) The event that **Samy** passes the statistics exam is **S**, and the event that **Ahmed** passes the same exam is **A**. The events **S** and **A** are independent and **P($\bar{s}$ and $\bar{A}$) = $\frac{1}{4}$, P(S) + P(A) = $\frac{23}{24}$** where $\bar{s}$ and $\bar{A}$ are the complements of events **S** and **A**, respectively.

Find the probability that both Samy and Ahmed pass the statistics exam.

(1.12) Two six-sided dice are rolled, and the number of spots turning up on each is observed. Consider the following two events**:**

A: The **sum** of the spots showing is **2**, **3**, or **12**.

B: The **sum** of the spots showing is an **even** number.

Are **A** and **B** independent events**?** Explain.

(Exam 2004)↓

(1.13) A large consumer goods company has been running a television advertisement for one of its soap products. A survey was conducted. On the basis of this survey, probabilities were assigned to the following events:

B = individual purchased the product

S = individual recalls seeing the advertisement

B and S = individual purchased the product and recalls seeing the advertisement

The probabilities assigned were**:**

P(B) = 0.2 , P(S) = 0.4 , and **P(B and S) = 0.12**.

(1) (What is the probability of an individual's purchasing the product given that the individual recalls seeing the advertisement**?**

(2) Find **P(B | S)**, where **S** is the complement of event **S**. Does seeing the advertisement increase the probability that the individual will purchase the product**?**

(3) Assume that individuals who do not purchase the company's soap product buy from its competitors. What would be your estimate of the company's market share**?**

(4) The company has also tested another advertisement and assigned its values of **P(S) = 0.3** and **P(B and S) = 0.1**. What is **P(B | S)** for this other advertisement**?** Which advertisement seems to be more effective than the other on consumer purchases**?**

(Exam 2004)

(1.14) Given**:**

$P(A_1) = 0.4$, $P(A_2) = 0.5$, $P(A_3) = 0.1$

$P(B \mid A_1) = 0.06$, $P(B \mid A_2) = 0.02$, $P(B \text{ and } A_3) = 0.08$

Find**:** $P(A_2 \mid B)$

(Exam 2005)↓

(1.15) Three machines **A**, **B**, and **C** produce the same identical units of a certain product. **Machine A** produces **40%**, **Machine B** produces **35%**, while **Machine C** produces **25%** of the total output of the product. All units produced by the three machines are mixed. From past experience it is known that **2%** of the units produced by **Machine A** are defective units, **4%** of the units produced by **Machine B** are defective units, while **5%** of the units produced by **Machine C** are defective units. A random unit of the machines' production is selected. Find**:**

(1) The probability that the selected unit is a **good** unit.

(2) If the selected unit was found to be a **defective** unit, find the probability that it was produced by either **Machine A** or **Machine B**.

(Exam 2005)

(1.16) Two players **A** and **B** play a series of independent games. For each game, the probability that **A** wins is **0.25**, the probability that **B** wins is **0.2** and, if neither player wins, the game is considered drawn.

For a single game,

(1) Calculate the following**:**

(a) The probability that the game is drawn.

(b) The probability that the game is either drawn or won by **A**.

(2) If two games are played, what is the probability that both are won by **B?**

(1.17) A fair die is cast; then **n** fair coins are tossed, where **n** is the number shown on the die. What is the probability of **exactly two** heads**?**

(Exam 2006)

(1.18) There 4 people being considered for the position of chief executive officer on an enterprise. 3 of the applicants are over 60 years of age. 2 are females, of which only one is over 60. Use the rule of multiplication to answer the following:

(1) What is the probability that a candidate is over 60 and female?

(2) Given that the person is over 60, what is the probability that the person is female?

(3) Given that the candidate is male, what is the probability that he is less than 60?

(Exam 2007)

(1.19) There are **four** defective items in a package of **10**. If **two** items are randomly selected one after another (**without replacement**), what is the probability of

(1) One defective and **one good** item being selected**?**

(2) Two good items being selected**?**

(1.20) A company is considering changing its starting hour from 8:00 A.M. to 7:30 A.M. A census of the company's **1,200** office and production workers shows **525** of its **750** production workers favor the change and a total of **840** workers favor the change. To further assess worker opinion, the region manager decides to talk with randomly selected workers.

(1) What is the probability a random selected worker will be in favor of the change**?**

(2) What is the probability a randomly selected worker will be **against** the change and be an **office** worker**?**

(3) Given that the selected worker was **in favor** of the change, find the probability that he is a **production** worker.

(4) Are the two events **"office worker"** and **"against" independent?** Explain. **(Exam 2008)**↓

(1.21) Given

- The two events **A** and **B** are independent.
- **P($\overline{A}$ and $\overline{B}$) = $\frac{1}{5}$ and P(A) + P(B) = $\frac{4}{5}$**

Show that **A** and **B** are **mutually exclusive** events.

(Exam 2008)

(1.22) Two **independent** events **A** and **B** are such that**:**

P(A) = 0.2 and **P(B) = 0.4**.

Find the following probabilities**:**

(1) P(**A and B**) **(2)** P(**A or B**)

(3) The probability of **A** occurring given that **B** has already occurred.

(1.23) A bag contains **20** balls, each of which is painted in two colors. **Five** are **red** and **white**, **seven** are **white** and **black**, and **eight** are **black** and **red**. A ball is chosen at random and seen to be partly **white**. What is the probability that its other color is **red?**

(1.24) A certain mass produced articles sometimes has a dimension defect and sometimes a surface defect. Let

$\mathbf{E_1}$ = an article has a dimension defect

$\mathbf{E_2}$ = an article has a surface defect

$\mathbf{E_3}$ = an article is non-defective

If **$P(E_1) = 0.06$** , **$P(E_2) = 0.07$** and **$P(E_3) = 0.9$**,

find **P**(article has **both** defects).

(Exam 2009)

(1.25) Given that, for two events **A** and **B**,

- **A and B are independent** - **P(A and B) = 0.4**
- **P(A) + P(B) = 0.6**

Show that **A** and **B** are **mutually exclusive** events.

(Exam 2013)↓

(1.26) A bag contains **3 red**, **4 white** and **5 black** balls. If **2** balls are taken, **without replacement**, what is the probability that they are all the **same color?**

(1.27) **Two** cards are drawn from a pack of playing cards without replacement. Find the probability that **at least** one is a **ten**. If only **one** card is drawn, what is the probability that this card is **5 given** that it was a **red** card**?**

(1.28) An unbiased **die F** has its faces numbered **1, 2, 3, 4, 5 and 6**. Another unbiased **die S** has its faces numbered **1, 1, 2, 2, 3 and 3**. In a game a card is selected at random from a pack of playing cards and if a **heart** is obtained die **F** is thrown; **otherwise** die **S** is thrown. Find the probability of obtaining **2**.

(Exam 2013)

(1.29) Two players **A** and **B** play a series of independent games. For each game, the probability that **A** wins is **0.3**, the probability that **B** wins is **0.2** and, if neither player wins, the game is considered drawn. For a single game,

(1) Calculate the following probabilities**:**

(a) The game is drawn.

(b) The game is either drawn or won by **A**.

(2) If two games are played, what is the probability that **both** are won by **B?**

(1.30) A bag contains **10** balls, each of which is painted in **two** colors. **Two** are **red** and **white**, **three** are **white** and **black**, and **five** are **black** and **red**. A ball is chosen at random and seen to be partly **red**. What is the probability that its other color is **white?**

(Exam 2014)↓

(1.31) The following table provides the marital status of a sample of 200 adults distributed by gender.

Gender	Marital Status				Total
	Single (S)	Married (R)	Widowed (W)	Divorced (D)	
Male (M)	10	50	12	8	80
Female (F)	15	75	18	12	120
Total	25	125	30	20	200

(1) An adult is selected at random, find the probability that the adult selected is**:**

(a) a **female**.

(b) divorced, **given** that the adult selected is a **male**.

(c) a **female or married**.

(2) On the basis of your results of **Part (1-b),** are the two events **"divorced"** and **"male" independent**? Explain.

(3) Are the events **"female"** and **"single" mutually exclusive**? Explain.

(Exam 2014)

(1.32) Two cards are chosen at random, **without replacement**, from a deck of **52** playing cards. Find the probability that

(1) Neither one is a **heart** **(2) At least one** is a **ten**.

(1.33) Two dice (D_1 and D_2) are rolled, find the **probability that** the **sum** of the two dice is **10**, **given that** the **first die** lands on **4**. That is, **$P(D_1 + D_2 = 10 \mid D_1 = 4)$**.

(1.34) Suppose the occurrence of **A** makes it more likely that **B** be will occur. In that case, show that the occurrence of **B** makes it more likely that **A** will occur. That is, show that, if **$P(B \mid A) > P(B)$**, then it is also true that **$P(A \mid B) > P(A)$**.

(Exam 2015)↓

(1.35) A fair die is thrown for as long as necessary for a **6** to turn up. Given that **6 does not turn up at the first throw**, what is the **probability** that **more than 4 throws** will be **necessary**?

(1.36) A real estate agent has a set of **3** keys, one of which will open the front door of a house he is trying to show to a client. If the keys are tried in completely random order, find the probability that:

(1) The **first** key **opens** the door. **(2) All 3** keys are **tried**.

(Exam 2015)

(1.37) Three men **A**, **B** and **C** share an office with a **single** telephone. Calls come in at random in the proportions **0.4** for **A**, **0.4** for **B**, **0.2** for **C**. Their work requires the men to **leave** the office at random times, so that **A** is out for **half** his working time and **B** and **C** each for a **quarter** of theirs. For calls arriving in working hours, find the probabilities that**:**

(1) No one is to **answer** the telephone.
(2) A call can be **answered** by the person being called.
(3) Three successive calls are for the **same** man.
(4) Three successive calls are for **different** men.
(5) A caller who **wants B** has to try **more than three** times to get him.

(1.38) In a group of **10** people, **3** support Party **X**, **5** support Party **Y** and **2** are **non-voters**. **Two** persons are randomly chosen **in succession** (without replacement).

(1) Given that the **first** person supports Party **X**, what is the probability that the **second** supports Party **Y**?
(2) What is the probability that **one** supporter of **each** party is drawn?

(Exam 2016) ↓

(3) What is the probability that **at least** one supporter of Party **Y** is drawn?

(1.39) Show that the probability of **A or B** can be written as**:**

$$P(A \text{ or } B) = P(A) + P(B)[1 - P(A \mid B)]$$

(1.40) In a campus restaurant, it was found that **40%** of all customers order **hot meals** and that **80%** of all customers are **students**. Further, **30%** of all customers who are s**tudents** order **hot meals**.

(1) What is the **probability** that a randomly chosen customer is **both** a **student** and orders a **hot meal**?

(2) If a randomly chosen customer orders a **hot meal**, what is the probability that the customer is a **student**?

(3) What is the probability that a randomly chosen customer **both does not** order a **hot meal and is not a student**?

Hint: For making it easier, set up a 2x2 table assuming that the total number of customers is 1000.

(4) Are the events **"customer is a student"** and **"customer orders a hot meal" independent**? **Explain**.

(Exam 2016)

(1.41) **Al-Ahly** and **Al-Zamalek** teams play a football match against each other. The probability for each team of scoring **0**, **1**, **2** goals are shown in the following table

(Hypothetical Data).

Number of Goals	0		1		2	
Team	Ahly	Zamalek	Ahly	Zamalek	Ahly	Zamalek
Probability of Scoring	0.1	0.3	0.6	0.5	0.3	0.2

(Exam 2017) ↓

(1) Find the **probabilities** of**:**

(a) Al-Ahly winning. **(b)** A draw **(No team wins)**.

(2) Based on your results of **Parts (a)** and **(b)**, find P(Al-Zamalek Winning).

(1.42) A girl has **three** unbiased (fair) dice. She starts a game by throwing one of the dice. If she **does not** throw a **6** the game **is over**, whereas if she **does throw** a **6** she then throws the **other two dice simultaneously (together)**. Calculate the probabilities that in one game she will have thrown

(1) Exactly **one 6**. **(2)** Exactly **two 6s**.

(1.43) Three events **A**, **B** and **C**. The events **A** and **C** are **mutually exclusive**. The events **A** and **B** are **independent**. Given that**:**

P(A) = 1/3 , **P(C) = 1/5** , **P(A or B) = 2/3**

(1) Find the probabilities**: (a)** P(A or C) **(b)** P(B) **(c)** P(A and B)

(2) Given also that**: P(B or C) = 3/5**, determine whether or not **B** and **C** are **independent**. **Justify** your answer.

(1.44) Box **(A)** contains **two red** balls and **three white** balls and **box (B)** contains **three red** balls and **one white** ball. A **card** is randomly drawn from a pack of playing cards and, if a **heart** shows, one ball is selected at random from **Box (A)**. Otherwise, a ball is randomly selected from **Box (B)**.

(1) Find the probability of selecting a **red** ball.

(2) The ball selected is **not replaced** and a **second** ball is selected at random from the **same** box. Find the probability that **both** balls are **white**.

(3) A ball is randomly withdrawn from **Box (A)** and **placed** in **Box (B)**. A ball is then taken at random from **Box (B)**, find the **probability** of getting a **red** ball.

(Exam 2017)

(1.45) The pages of a book are numbered from **1** to **200**. If a page is chosen at random, what is the probability that its number will contain just **two** digits?

(1.46) Suppose that A_1, A_2, and **B** are events where A_1 and A_2 are **mutually exclusive** events and
$P(A_1) = 0.8$, $P(A_2) = 0.2$, $P(B \mid A_1) = 0.1$, $P(B \mid A_2) = 0.3$
Use these informations to find**: (1)** $P(A_1 \mid B)$ **(2)** $P(A_2 \mid B)$

(1.47) A restaurant has the following data on the age and marital status of **100** customers**:**

		Marital Status	
		Single (S)	**Married (M)**
Age	**Under 30 (U)**	**14**	**6**
	30 or over (V)	**16**	**64**

(1) What is the **probability** of finding a customer who is **married and under the age of 30**?
(2) If a customer is **30 or over**, what is the **probability** that he or she is **single**?
(3) Is **marital status independent** of **age**? **Explain**, using probabilities.

(1.48) On a small island, **2** rabbits are caught. Each is tagged (marked) so that it can be recognized again and is then released. Next day **5** more (untagged) rabbits are caught. These too are tagged and released. On the third day, **4** rabbits are caught and **two** of them are found to be already **tagged**.
Assuming the rabbit population is constant and that tagged and untagged rabbits are equally likely to be caught, find

(Exam 2018) ↓

(1) the **smallest number** of rabbits there can be **on this island**.

(2) the **probability** of the **second day's catch** if there are exactly **12** rabbits on the island.

(1.49) A bag contains **two black** balls and **three red** balls. A ball is drawn and **replaced** by a ball of the **opposite** color. Then another ball is drawn from the bag. Find the probability that

(1) the **first** ball drawn was **red**, **given** that the **second** ball drawn was **black**.

(2) the **first** ball drawn was **black**, given that the **second** ball drawn was **black**.

Hint: Use your results of **Part (1)**.

(Exam 2018)

(1.50) If the occurrence of one event means that another can't happen, then the events are

(A) Independent **(B) Mutually exclusive**

(C) Empirical **(D) None of the above**

(1.51) The statement that **"P(A | B) = P(B | A)"** whenever **A** and **B** are **independent** events is

(A) Always true **(B) Never true**

(C) Not enough information; we would need to know if A and B are mutually exclusive events

(D) Not enough information; we need to know if the events are equally likely

Two dice (A and B) are rolled.
Answer the following two Questions: (1.53) and (1.54)

(1.52) What is the **probability: P(A + B = 8 | A = even number)**?

(A) 5/18 **(B) 1/6** **(C) 5/6** **(D) 4/9**

(Exam 2019) ↓

(1.53) The **probability** that **(at least)** one of the dice is a **4 or** the **sum of the two dice equal to 7** is

(A) 1/9 (B) 4/9 (C) 5/12 (D) 7/12

If a **card** is **chosen** from a pack of **playing cards**.
Answer the following <u>two</u> Questions: (1.55) and (1.56)

(1.54) What is the **probability** of getting a **10**?

(A) 10/52 (B) 4/52 (C) 1/52 (D) 2/13

(1.55) The **probability** that this card is **a heart given** that it was **red** is **(A) 0.5 (B) 0.3 (C) 0.6 (D) 0.4**

(1.56) A bag contains **one red**, **2 white** and **2 black** balls. If **two** balls are taken **(without replacement)**, what is the **probability** that they are of **different** colors?

(A) 0.6 (B) 0.4 (C) 0.8 (D) 0.2

(1.57) Suppose you are taking a **multiple-choice** test with **4** choices for each question, one of which is correct. In answering a question on this test, the **probability** that you **know** the answer is **0.6**. If you **don't know** the answer, you choose one **at random**. What is the **probability** that you **knew the answer** to a question, **given** that you **answered** it **correctly**?

(A) 0.82 (B) 0.84 (C) 0.88 (D) 0.86

(1.58) A certain product has **two** defects (**D1** and **D2**). Let

D1 = The product has defect 1,
D2 = The product has Defect 2
G = The product is non-defective

If **P(D1) = 0.05 , P(D2) = 0.07 and P(G) = 0.8**

Find **P(**the product has **both** defects**)**.

(A) 0.08 (B) 0.04 (C) 0.06 (D) 0.12

(Exam 2019)

Chapter (2)
Discrete Probability Distributions

Contents

Chapter 2
Discrete Probability Distributions

2.1 Probability Distribution

Definition:

Discrete Variable:

A variable whose values are countable. The discrete variable can assume only certain values with no intermediate values.

For example, the number of cars sold on any given day at a car dealership is a discrete variable because the number of cars sold must be 0, 1, 2, 3, . . . and we can count it. The number of cars sold cannot be between 0 and 1, or between 1 and 2. Other examples of discrete variables are the number of people visiting a bank on any day, the number of cars in a parking lot, the number of cattle owned by a farmer, and the number of students in a class.

Probability Distributions:

The probability distribution of a discrete random variable X is a list of each possible value of X together with the probability that X takes that value in one trial of the experiment.

The probabilities in the probability distribution of a random variable *X* must satisfy the following two conditions**:**

- Each probability P(x) must be between 0 and 1**:** $0 \le P(x) \le 1$.
- The sum of all the probabilities is 1; $\Sigma P(x) = 1$.

2.2 Discrete Probability Distribution:

Example (2.1):

Each of the following tables lists certain values of x and their probabilities. Determine whether or not each table represents a valid probability distribution.

(a)		(b)		(c)	
x	P(x)	x	P(x)	x	P(x)
0	0.16	1	0.24	2	0.35
1	0.22	2	0.36	3	0.45
2	0.36	3	0.25	4	0.52
3	0.18	4	0.15	5	-0.32

Solution:

(a) Because each probability listed in this table is in the range 0 to 1, it satisfies the first condition of a probability distribution. However, the sum of all probabilities is not equal to 1.0. because $\Sigma P(x) = 0.16 + 0.22 + 0.36 + 0.18 = 0.92$. Therefore, the second condition is not satisfied.

Consequently, this table does not represent a valid probability distribution.

(b) Each probability listed in this table is in the range 0 to 1. Also, $\Sigma P(x) = 0.24 + 0.36 + 0.25 + 0.15 = 1.0$. Consequently, this table represents a valid probability distribution.

(c) Although the sum of all probabilities listed in this table is equal to 1.0, one of the probabilities is negative. This violates the first condition of a probability distribution. Therefore, this table does not represent a valid probability distribution.

Example (2.2):

A fair coin is tossed twice. Let X be the number of heads that are observed.

a. Construct the probability distribution of X.

b. Find the probability that at least one head is observed.

Solution:

Referring to Example (1.3) in Chapter (1), Page 7

a. The possible values that X can take are 0, 1, and 2. Each

of these numbers corresponds to an event of equally likely outcomes for this experiment: X = 0 for (TT) , X = 1 for (HT,TH), and X = 2 for (HH). The probability of each of these events, hence of the corresponding value of X, can be found simply by counting, to give the following distribution:

x	0	1	2
P(x)	0.25	0.50	0.25

This table is the probability distribution of X.

b. "At least one head" is the event $X \geq 1$, which is the mutually exclusive events $X = 1$ and $X = 2$. Thus

$P(X \geq 1) = P(1) + P(2) = 0.5 + 0.25 = 0.75$

Example (2.3):

A pair of fair dice is rolled, one white and one green. Let X denote the sum of the number of dots on the top faces.
Construct the probability distribution of X.

Solution:

We use notation like (1,2) to mean "1 on the white, 2 on the green". With this, the equally likely outcomes are

(1,1)	**(1,2)**	**(1,3)**	**(1,4)**	**(1,5)**	**(1,6)**
(2,1)	**(2,2)**	**(2,3)**	**(2,4)**	**(2,5)**	**(2,6)**
(3,1)	**(3,2)**	**(3,3)**	**(3,4)**	**(3,5)**	**(3,6)**
(4,1)	**(4,2)**	**(4,3)**	**(4,4)**	**(4,5)**	**(4,6)**
(5,1)	**(5,2)**	**(5,3)**	**(5,4)**	**(5,5)**	**(5,6)**
(6,1)	**(6,2)**	**(6,3)**	**(6,4)**	**(6,5)**	**(6,6)**

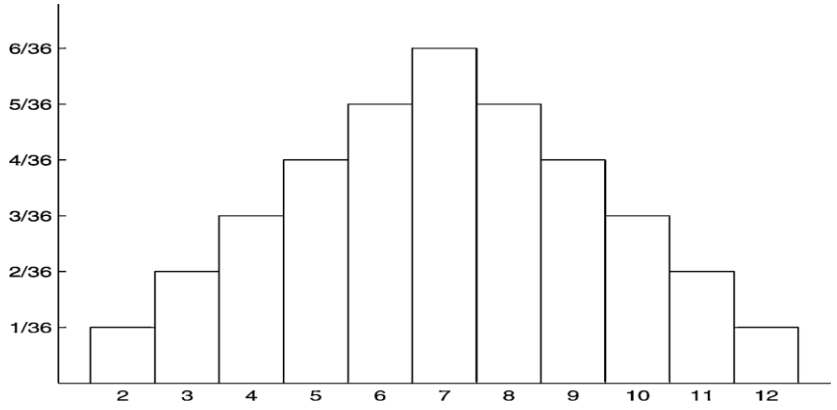
The values of x are 2, 3, 4, up to 12 (starting from 1 + 1 = 2 to 6 + 6 =12). The probability of each of these events, hence of the

corresponding value of X, can be found simply by counting, to give the following probability distribution**:**

Probability Distribution for Tossing Two Fair Dice

X	2	3	4	5	6	7	8	9	10	11	12
P(x)	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

A histogram that graphically illustrates the probability distribution is given below. From this histogram, it is obviously noted that this histogram is symmetric around the value 7 (A + B = 7) which is the mean of the distribution.



Definition:

The Mean of The Probability Distribution of a Discrete Random Variable:

The mean (also called the expected value) of a discrete random variable X is

$$E(X) = \mu = \Sigma xP(x)$$

The mean **μ** of a discrete random variable X is a number that indicates the average value of X over numerous trials of the experiment.

Definition:

The Variance σ^2 of The Probability Distribution of A Discrete Random Variable X:

They are numbers that indicate the variability of X over numerous trials of the experiment.

The **variance** of a discrete random variable X is given by

$$\textbf{Var (x)} = \sigma^2 = \Sigma(x - \mu)^2 P(x)$$

Which by algebra is equivalent to the formula:

$$\textbf{Var (x)} = \sigma^2 = \{\Sigma x^2 P(x)\} - \{E(x)\}^2$$

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

The **standard deviation** = $\sqrt{\mathbf{Var(x)}} = \sigma$

Example (2.4):

Let x be the number of errors that appear on a randomly selected page of a book. The following table lists the probability distribution of x.

X	0	1	2	3
P(x)	0.75	0.15	0.08	0.02

(1) Find the following probabilities**:**

(a) P(x > 1) **(b)** $P(1 \leq x < 3)$

(2) Find the mean and standard deviation of X.

Solution:

(1) (a) P(x > 1) = P(2) + P(3) = 0.08 + 0.02 = 0.1

(b) $P(1 \leq x < 3)$ = P(1) + P(2) = 0.15 + 0.08 = 0.23

(2) $E(X) = \mu = \Sigma x P(x) = 0 \times 0.75 + 1 \times 0.15 + 2 \times 0.08 + 3 \times 0.02 = 0.37$

$E(x^2) = (0)^2(0.75) + (1)^2(0.15) + (2)^2(0.08) + (3)^2(0.02) = 0.65$

$\text{Var}(x) = 0.65 - (0.37)^2 = 0.5131$

Example (2.5):

For the discrete random variable X in **Example (2)**, find

(a) The **mean** and **variance** of the variable x.

(b) The **mean** and **standard deviation** of the variable **y,** where **y = 2x + 3**

Solution:

(a) The formulas in the definitions give

$E(x) = \mu = \Sigma xP(x) = 0 \times 0.25 + 1 \times 0.5 + 2 \times 0.25 = 1$

$E(x^2) = \Sigma x^2P(x) = (0)^2(0.25) + (1)^2(0.5) + (2)^2(0.25) = 1.5$

$\text{Var}(x) = \sigma^2 = 1.5 - (1)^2 = 0.5$

(b) Since $y = 2x + 3$;

$E(y) = 2E(x) + 3 = 2(1) + 3 = 5$

$\text{Var}(y) = 4[\text{Var}(x)] = 4(0.5) = 2$

Thus, $S(y) = \sqrt{2} = 1.414$

Hint: The value of the variance of a data set is not changed by adding to (or subtracting from) the same constant to each value of data set. This is not the situation in case of multiplication or division.

We also can find the standard deviation of y by constructing the probability distribution of y as follows.

The probability distribution of y is:

y = 2x + 3	3	5	7
P(y)	0.25	0.50	0.25

Hint: P(x = 0) = P(y = 3) , P(x = 1) = P(y = 5) and P(x = 2) = P(y = 7)

From this probability distribution, we obtain:

$E(y) = \Sigma yP(y) = 3(0.25) + 5(0.5) + 7(0.25) = 5$

$E(y^2) = \Sigma y^2P(y) = (3)^2(0.25) + (5)^2(0.5) + (7)^2(0.25) = 27$

Therefore, $Var(y) = 27 - (5)^2 = 2$, and

$$S(y) = \sqrt{2} = 1.414$$

The same results obtained before.

Obviously, the first method is easier than the second one. as it requires fewer calculations, especially if these measures are also required for the variable x.

Example (2.6):

Check whether the function given by

$$P(x) = \frac{x + 2}{25} \text{ for } x = 1, 2, 3, 4, 5$$

can serve as the probability distribution of a discrete random variable.

Solution:

Substituting the different values of x, we get P(1) = 3/25 , P(2) = 4/25 , P(3) = 5/25, , P(4) = 6/25 , and P(5) = 7/25.

Since P(1) + P(2) + P(3) + P(4) + P(5) = 3/25 + 4/25 + 5/25 + 6/25 + 7/25 = 1

Thus, the given function can serve as the probability distribution of a random variable having the range {1, 2, 3, 4, 5}.

Example (2.7):

Let X be a discrete random variable with the following probability distribution**:**

x	0	1	2	3
P(x)	A	0.4	B	C

Given: E(x) = 1.6 and Var (x) = 9.84, find the values of A, B, and C.

Solution:

Since, $E(x) = 0 \times A + 1 \times 0.4 + 2 \times B + 3 \times C = 1.6$

Which gives **$2B + 3C = 1.2$** **(1)**

$\text{Var}(x) = E(x^2) - \{E(x)\}^2 = 0.84$

$= E(x^2) - (1.6)^2 = 0.84$

So, $E(x^2) = 3.4 = \Sigma x^2 P(x) = (0)^2(a) + (1)^2(0.4) + (2)^2(b) + (3)^2(0.2)$

This gives **$4B + 9C = 3$** **(2)**

By solving Equations (1) and (2), we get**:** $B = 0.3$ and $C = 0.2$

Since $\Sigma P(x) = 1$

Then $A + B + C = 1 - 0.4 = 0.6$

which gives $A = 0.1$ ($B = 0.3$ and $C = 0.2$)

Exercises for Section 2.2
(Exam Questions)

(2.1) Let x denote the number of accidents that occur in a city during a week. The following table lists the probability distribution of x.

x	0	1	2	3
p(x)	0.2	0.4	0.2	0.1

(a) Determine the probability that the number of accidents that will occur during a given week in this city is**:**

(i) exactly 3 **(ii)** at least 2 **(iii)** less than 2 **(iv)** one to 3

(b) Find E(x) and Var(x)

(c) Is E(x) a possible value of x?

(Exam 1998)

(2.2) Let x be the number of errors that a random selected page of a book contains. The following table lists the probability distribution of x:

x	0	1	2	3	4
P(x)	0.1	0.2	0.4	0.2	0.1

(1) Find the mean and variance of y, where y = 2x

(2) Comment on the symmetry or skewness of the distribution.

(3) Find the following probabilities:

(a) P(x $\leq$ 3) **(b) P(x = 2)** **(c) P(1 $\leq$ x $\leq$ 4)**

(Exam 1999)

(2.3) A random variable can assume only even integer values between 1 and 9. It is distributed in such a way that $P(x) = \frac{x}{\sum x}$.
Find the following:

(a) The probability distribution of X.

(b) $P(x \geq 6)$ **(c)** E(X)

(Exam 2000)

(2.4) Let **x** represent the number of children under 18 years old in an Egyptian family. The probability distribution of **x** is

x	**0**	**1**	**2**	**3**	**4**	**5**
P(x)	**0.50**	**0.21**	**0.19**	**0.07**	**0.02**	**0.01**

(a) Comment on the symmetry or skewness of the distribution.

(b) What is the most likely number of children in a family**?**

(c) Find the following**:**

(i) $P(x = 3)$ **(ii)** $P(x \geq 4)$ **(iii)** $P(2 < x \geq 5)$

(iv) E(2x - 1) **(v)** Var (2x + 1)

(Exam 2002)

(2.5) Let x be a random variable with the following probability distribution**:**

x	**1**	**2**	**3**
p(x)	**0.25**	**0.25**	**0.50**

(a) Comment on the symmetry or skewness of the distribution.

(b) Find the probability that $\mathbf{0 \leq X \leq 2}$.

(c) Find the expected value and variance of **Y = 2X – 3**.

(Exam 2003 Qena)

(2.6) Three tables listed below show "random variables" and their "probabilities". However, only one of these is actually a probability distribution.

Table (1)		**Table (2)**		**Table (3)**	
X	P(x)	X	P(x)	X	P(x)
0	0.4	0	0.4	0	0.1
1	0.3	1	0.3	1	0.5
2	0.3	2	0.2	2	0.2
3	0.1	3	0.1	3	0.1

(1) Which is it? Justify your answer.

(2) Find the probability that x is:

(a) Exactly 3 **(b)** No more than 1 **(c)** $P(1 < x \leq 3)$

(3) Find the mean and variance for the probability distribution.

(4) Suppose that **y = 2x + 1**. For each value of x, determine the value of y. What is the **probability distribution of y**?

(5) Calculate the mean, variance, and standard deviation of y.

(6) Use the laws of expected value and variance to calculate the mean, variance, and standard deviation of y from the mean, variance, and standard deviation of x. Compare your answers in **Parts (5)** and **(6)**. Are they the same?

(Exam 2007)

(2.7) Let **x** denote the number of accidents that occur in a city during a week. The following table presents the probability distribution of **x**.

x	1	2	3
P(x)	0.2	0.5	0.3

(Exam 2013)↓

(1) Find the following probabilities**:**

(a) P(x = 2) (b) P(x < 3) (c) P(x = at most 2)

(2) Calculate the **mean** and **standard deviation** of **x**.

(3) Find the **mean** and **variance** of **y**, where **y = 2x + 3**.

(Exam 2013)

(2.8) A factory manager collected data on the number of equipment breakdowns per day. From those data, he derived the probability distribution shown in the following table, where **x** denotes the **number of breakdowns** on a given **day**.

x	0	1	2
P(x)	0.80	0.15	0.05

(1) Find the following probabilities**:**

(a) P(x = 1) (b) P(2 ≤ x < 3) (c) P(x = at most 2)

(2) Determine the **mean** and **standard deviation** of **x**.

(3) Find the **mean** and **variance** of **Y**, where **Y = 10 – 2x**.

(4) On **average**, how many breakdowns occur **per day**?

(5) About how many **breakdowns** are expected during a **1- month** period, assuming **20** work days per month?

(Exam 2015)

(2.9) Let ***x*** denote the number of accidents that occur in a city during a day. The following table presents the **probability distribution** of **x**.

Number of Accidents (x)	**0**	**1**	**2**
Probability: P(x)	**0.2**	**0.7**	**0.1**

(1) Is this a **valid** probability distribution? **Explain**.

(Exam 2017)↓

(2) Find the following probabilities**:**

(a) P(x = 1) **(b) P(0 < x ≤ 1)**

(3) Find the **mean** and **variance** of **Y**, where **Y = 3x + 2**.

(4) About **how many accidents** are expected during a **50- day** period?

(Exam 2017)

(2.10) Refer to the discrete **probability distribution** provided in the table below**:**

x	0	1	2	3
P(x)	0.05	0.15	0.60	?

(1) The **probability: P(1 ≤ x < 3)** is

(A) 0.80 **(B) 0.20** **(C) 0.75** **(D) 0.65**

(2) If **y = 4x - 2**, what is the **variance** of **y?**

Hint: No rounding required.

(A) 9.24 **(B) 8.76** **(C) 6.25** **(D) 7.85**

(MCQ Exam 2019)

2.3 Binomial Distribution

What is Binomial Distribution?

Binomial distribution is a common probability distribution that models the probability of obtaining one of two outcomes under a given number of parameters. It summarizes the number of trials when each trial has the same chance of attaining one specific outcome.

The Binomial distribution is a discrete probability distribution of the successes in a sequence of n independent experiments, each of which yields success with probability p and failure with probability q, where q = 1 - p.

Criteria of Binomial Distribution:

Binomial distribution models the probability of occurrence of an event when specific criteria are met. Binomial distribution involves the following rules that must be present in the process in order to use the Binomial probability formula:

1. Fixed Trials

The process under investigation must have a fixed number of trials that cannot be altered in the course of the analysis. During the analysis, each trial must two outcomes, although each trial may yield a different outcome.

In the Binomial probability formula, the number of trials is represented by the letter “n”. An example of independent trials may be tossing a coin or rolling a die. When tossing a coin, the first event is independent of the subsequent events. The number of times that each trial is conducted is known from the start. If a coin is tossed 10 times, each toss of the coin is a trial.

2. Independent Trials

The other condition of a Binomial probability is that the trials are

independent of each other. In simple terms, the outcome of one trial should not affect the outcome of the subsequent trials.

When using certain sampling methods, there is a possibility of having trials that are not completely independent of each other, and Binomial distribution may only be used when the size of the population is large compared with the sample size.

3. Fixed Probability of Success

In a Binomial distribution, the probability of getting a success must remain the same for the trials we are investigating. For example, when tossing a coin, the probability of flipping a coin is 0.5 for every trial we conduct, since there are only two possible outcomes.

In some sampling techniques, such as sampling without replacement, the probability of success from each trial may vary from one trial to the other. For example, assume that there are 50 boys in a population of 100 students. The probability of picking a boy from that population is 0.5.
In the next trial, there will be 49 boys out of 99 students. The probability of picking a boy in the next trial is 0.49. It shows that in subsequent trials, the probability from one trial to the next will vary slightly from the prior trial.

4. Two mutually exclusive outcomes

In Binomial probability, there are only two mutually exclusive outcomes, i.e., success or failure. While success is generally a positive term, it can be used to mean that the outcome of the trial agrees with what you have defined as a success, whether it is a positive or negative outcome.
For example, when a business receives a consignment of lamps with a lot of breakages, the business can define success for the

trial to be every lamp that has broken glass. A failure can be defined as when the lamps have zero broken glasses.

In our example, the instances of broken lamps may be used to denote success as a way of showing that a high proportion of the lamps in the consignment is broken. and that there is a low probability of getting a consignment of lamps with zero breakages.

In general, if the random variable X follows the Binomial distribution with the two parameters n and p. That is,

$$\mathbf{X \sim B(n, p)}$$

The probability of getting exactly x successes in n trials is given by the probability function:

$$\mathbf{P(x) = \binom{n}{x} p^x q^{n-x}},$$

$$\mathbf{x = 0, 1, 2, 3, \ldots\ldots\ldots, n}$$

Where

n = Total number of trials

x = Number of successes in the n trials

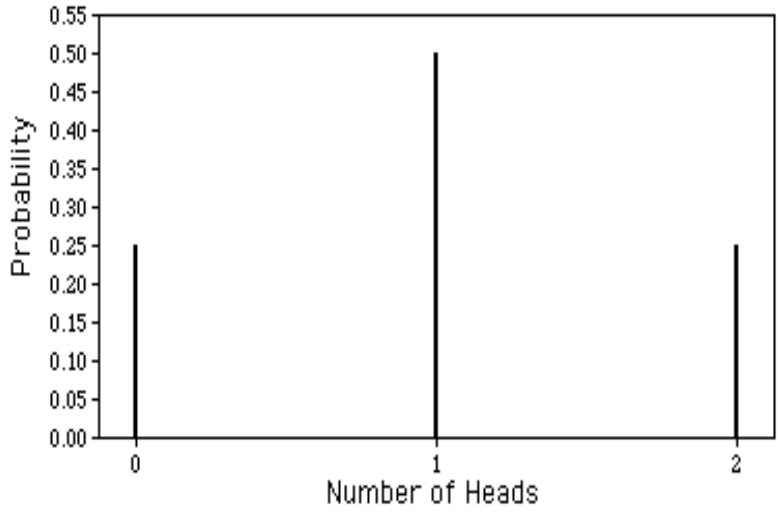
n - x = Number of failures in n trials

P = the probability of success on each trial

q = 1 − p the probability of failure

$\binom{\mathbf{n}}{\mathbf{x}} = \frac{\mathbf{n!}}{\mathbf{x!(n-x)!}}$ is the **Binomial coefficient** (hence the name of the distribution) **"n choose k"** also denoted by **C(n,x) or $^{n}C_{x}$**. They are simply written by the formula $\binom{\mathbf{n}}{\mathbf{x}}$.

This is a graphic representation of a Binomial probability distribution with p = q = 0.5.



The Binomial distribution is frequently used to model the number of successes in a sample of size n drawn with replacement from a population of size N. If the sampling is carried out without replacement, the draws are not independent and so the resulting distribution is not a Binomial one. In this case, the hypergeometric distribution is used. We will discuss this distribution later.

Examples of Binomial Distribution:

Example (2.8): Tossing A Coin Twice

The following table presents the outcomes of tossing a coin twice.

Outcome	First Toss	Second Toss	Both
1	Head (H)	Head (H)	HH
2	Head (H)	Tail (T)	HT
3	Tail (T)	Head (H)	TH
4	Tail (T)	Tail (T)	TT

From this table, the following probabilities can be obtained:

(1) P(Two Heads) = P(HH) = 1/4

(2) P(One Head) = P(HT) or P(TH) = 2/4

(3) P(Two Tails) = P(TT) = 1/4

Note that: P(H) = P(T) = 0.5

Alternatively, we can apply the information in the Binomial probability formula of the Binomial distribution as follow**:**
The first step in finding the Binomial probability is to verify that the situation satisfies the four rules of Binomial distribution:

- Number of fixed trials (n): 2 (Tossing a coin twice)
- Number of mutually exclusive outcomes: 2 (Head and Tail)
- The probability of success (p): 0.5
- Independent trials: Yes

$$\mathbf{P(x = x) = P(x) = \binom{n}{x} p^x q^{n-x}}$$

Where

$$\binom{\mathbf{n}}{\mathbf{x}} = \frac{\mathbf{n!}}{\mathbf{x!\,(n-x)!}}$$

(1) x = 2 and n = 2

P = Probability of (success) = Probability of getting a Head =0.5

$$P(HH) = P(x = 2) = \binom{n}{x} p^x q^{n-x}$$

$$= \binom{2}{2} (0.5)^2 (0.5)^0 = 0.25$$

(2) x = 1 and n = 2

P = Probability of (success) = Probability of getting a Tail = 0.5

$$P(\text{One T}) = \binom{2}{1} (0.5)^1 (0.5)^1 = 0.5$$

(3) x = 2 and n = 2

P = Probability of (success) = Probability of getting a Tail = 0.5

$$P(TT) = \binom{2}{2} (0.5)^2 (0.5)^0 = 0.25$$

Mean, Standard Deviation, and Variance of the Binomial Distribution:

- **Variance:** a measure of how far a set of numbers is spread out.
- **Mean:** one measure of the central tendency either of a probability distribution or of the random variable characterized by that distribution.
- **Standard deviation:** shows how much variation or dispersion exists from the average (mean), or expected value.

As with most probability distributions, examining the different properties of Binomial distributions is important to truly understanding the implications of them. The mean, variance, and standard deviation are three of the most useful and informative properties to explore. We'll take a look at these different properties and how they are helpful in establishing the usefulness of statistical distributions. The easiest way to understand the mean, variance, and standard deviation of the Binomial distribution is to use a real-life example.

Consider a coin-tossing experiment in which you tossed a coin 12 times and recorded the number of heads. If you performed this experiment over and over again, what would the mean number of heads be? On average, you would expect half the coin tosses to come up heads. Therefore, the mean number of heads would be 6. In general, the **mean** of a Binomial distribution with parameters **n** (the number of trials) and **p** (the probability of success for each trial) is**:**

$$E(x) = np$$

Where **E(x)** is the **mean** of the **Binomial distribution.**

The **standard deviation (s)** and **variance (s^2)** of the Binomial distribution are:

$$\mathbf{S = \sqrt{np(1-p)}}$$

$$\mathbf{Var\,(x) = s^2 = np(1 - p) = npq},$$

where **s** and $\mathbf{s^2}$ are the standard deviation and **variance** of the Binomial distribution, respectively.

The coin was tossed 12 times, so n = 12. A coin has a probability of 0.5 of coming up heads. Therefore, the mean and standard deviation can therefore be computed as follows:

$E(x) = np = 12(0.5) = 6$

$Var\,(x) = np(1 - p) = 12(0.5)(1 - 0.5) = 3$

Naturally, the standard deviation (s) is the square root of the variance (s^2). That is, $S = \sqrt{3} = 1.73$

Example (2.9):

Suppose, according to the latest police reports, **80%** of all crimes are unresolved, and in your town, at least three of such crimes are committed. The three crimes are all independent of each other. From the given data.

(1) what is the probability that one of the **three** crimes will be **solved**?

(2) Find the probability that **one** of the three crimes will be **unsolved**.

Solution:

(1) We find the probability that one of the crimes will be solved in the three independent trials. It is shown as follows:

Trial 1 = Solved 1st, unsolved 2nd, and unsolved 3rd

= 0.2 x 0. 8 x 0.8 = 0.128

Trial 2 = Unsolved 1st, solved 2nd, and unsolved 3rd

= 0.8 x 0.2 x 0.8 = 0.128

Trial 3 = Unsolved 1st, unsolved 2nd, and solved 3rd

= 0.8 x 0.8 x 0.2 = 0.128

Total (for the three trials):

= 0.128 + 0.128 + 0.128 = 0.384

Alternatively, we can apply the information in the Binomial probability formula, as follows:

The first step in finding the Binomial probability is to verify that the situation satisfies the four rules of Binomial distribution:

- Number of fixed trials (n): 3 (Number of crimes)
- Number of mutually exclusive outcomes: 2 (solved and unsolved)
- The probability of success (p): 0.2 (20% of cases are solved)
- Independent trials: Yes

Applying the Binomial distribution gives: $\mathbf{x}$ = 1 and n = 3

In this case, success is to solve the crime. So, p = 0.2

P(1) = P(x = 1) = $\binom{3}{1}(0.2)^1(0.8)^2$ = 0.384

the same result obtained before.

(2) Trial 1 = Unsolved 1st, Solved 2nd, and Solved 3rd

= 0.8 x 0. 2 x 0.2 = 0.032

Trial 2 = Solved 1st, Unsolved 2nd, and Solved 3rd

= 0.2 x 0.8 x 0.2 = 0.032

Trial 3 = Solved 1st, Solved 2nd, and Unsolved 3rd

= 0.2 x 0.2 x 0.8 = 0.032

Total (for the three trials):

= 0.032 + 0.032 + 0.032 = 0.096

Alternatively, applying the Binomial formula gives**:**

n = 3 , x = 1 , p = 0.8

P(One Crime Unsolved) = P(1) = $\binom{3}{1}(0.8)^1(0.2)^2$ = 0.096

The same result obtained before.

Hint: Regarding the following examples, we will suffice to solve them using only the Binomial Distribution, and all of these examples meet the conditions necessary to apply this distribution.

Example (2.10):

A multiple-choice test has **5** questions and each question has **4** choices, one of which is correct. You do not have a clue about the subject matter and guess your way through the test. What is the probability that you guess?

(A) Exactly two questions correctly?

(B) Exactly one question incorrectly?

(C) At least one question correctly?

Solution:

(A) Success is to guess the answer correctly, so

n = 5 , x = 2 , p = $\frac{1}{4}$, q = $\frac{3}{4}$

P(1) = $\binom{5}{2}(\frac{1}{4})^2\left(\frac{3}{4}\right)^3$ = 0.2637

(B) Success is to guess the answer incorrectly, hence

n = 5 , x = 1 , p = $\frac{3}{4}$, q = $\frac{1}{4}$

P(1) = $\binom{5}{1}(\frac{3}{4})^1\left(\frac{1}{4}\right)^4$ = 0.0586

(c) Success is to guess the answer correctly, therefore

n = 5 , x ≥ 1 , p = $\frac{1}{4}$, q = $\frac{3}{4}$

$$P(X \geq 1) = 1 - P(0) = 1 - \binom{5}{0} \left(\frac{1}{4}\right)^0 \left(\frac{3}{4}\right)^5$$

$$= 1 - 0.2373 = 0.7627$$

Example (2.11):

A die is tossed **3** times. What is the probability of

(1) No fives turning up? **(2)** **3** fours turning up?

Solution:

(1) Success is to get a number other than 5;

$n = 3$, $x = 3$, $P(5) = \frac{5}{6}$

$$P(0) = \binom{3}{3} \left(\frac{5}{6}\right)^3 \left(\frac{1}{6}\right)^0 = 0.5787$$

(2) Success is to get a 4, so

$n = 3$, $x = 3$, $P(4) = \frac{1}{6}$

$$P(3) = \binom{3}{3} \left(\frac{1}{6}\right)^3 \left(\frac{5}{6}\right)^3 = 0.0965$$

Example (2.12):

A coin is tossed four times. Calculate the probability of obtaining more heads than tails.

Solution:

To obtain more heads than tails, the number of heads must be at least 3.

Success is to get a head;

$n = 4$, $x = 3$ or 4 , $P(4) = \frac{1}{2}$

$P(x \geq 3) = P(x = 3) + P(x = 4)$

$$= \binom{4}{3} (0.5)^3 (0.5)^1 + \binom{4}{4} (0.5)^4 (0.5)^0 = 0.3125$$

Example (2.13):

An agent sells life insurance policies to five equally aged, healthy people. According to recent data, the probability of a person living in these conditions for 30 years or more is 2/3. Calculate the probability that after 30 years:

(1) All five people are still living.

(2) At least three people are still living.

(3) Exactly two people are still living.

Solution:

(1) All five people are still living.

$X \sim B(5, \frac{2}{3})$, $p = \frac{2}{3}$, $q = \frac{1}{3}$

$P(x = 5) = \binom{5}{5}\left(\frac{2}{3}\right)^5\left(\frac{1}{3}\right)^0 = 0.132$

(2) At least three people are still living.

$P(x \geq 3) = P(x = 3) + P(x = 4) + P(x = 5)$

$= \binom{5}{3}\left(\frac{2}{3}\right)^3\left(\frac{1}{3}\right)^2 + \binom{5}{4}\left(\frac{2}{3}\right)^4\left(\frac{1}{3}\right)^1 + \binom{5}{5}\left(\frac{2}{3}\right)^5\left(\frac{1}{3}\right)^0$

$= 0.791$

(3) Exactly two people are still living**:**

$n = 5$, $x = 2$, $p = \frac{2}{3}$

$P(2) = \binom{5}{2}\left(\frac{2}{3}\right)^2\left(\frac{1}{3}\right)^3 = 0.164$

Example (2.14):

The probability of a man hitting the target at a shooting range is 1/4. If he shoots 10 times.

(a) What is the probability that he hits the target exactly three times?

(b) What is the probability that he hits the target at least once?

Solution:

(a) Success is to hit the target ;

$X \sim B(10 , \frac{1}{4})$, x = 3 , p = 1/4 , q = 3/4

$P(x = 3) = \binom{10}{3}\left(\frac{1}{4}\right)^3\left(\frac{3}{4}\right)^7 = 0.25$

P(at least once) = $P(x \geq 1) = 1 - P(0)$

$= 1 - \binom{10}{0}\left(\frac{1}{4}\right)^0\left(\frac{3}{4}\right)^{10} = 0.9437$

Example (2.15):

There are 10 red and 15 blue balls in a box. A ball is chosen at random and it is noted whether it is red. The process repeats, returning the ball 10 times. Calculate the expected value and the standard deviation of this game.

Solution:

P(Red Ball) = $\frac{10}{10 + 15} = 0.4$

X ~ B(10 , 0.4)

E(x) = np = 10(0.4) = 4

Var (x) = npq = 10(0.4)(0.6) = 2.4

Standard deviation (s) = $\sqrt{2.4} = 1.55$

Example (2.16):

A pharmaceutical lab states that a drug causes negative side effects in **3** of every **100** patients. To confirm this affirmation, another laboratory chooses **5** people at random who have

consumed the drug. What is the probability of the following events?

(1) None of the five patients experience side effects.

(2) At least two experience side effects.

(3) What is the average number of patients that the laboratory should expect to experience side effects if they choose 100 patients at random?

Solution:

P(Side Effect) = $\frac{3}{100}$ = 0.03

(1) None of the five patients experience side effects.

X ~ B(100 , 0.03)

$$P(x = 0) = \binom{5}{0}(0.03)^0(0.97)^5 = 0.8687$$

(2) At least two experience side effects.

$P(X \geq 2) = 1 - P(x < 2) = 1 - [\, P(x = 0) + P(x = 1)]$

$$= 1 - \{\binom{5}{0}(0.03)^0(0.97)^5 + \binom{5}{1}(0.03)^1(0.97)^4\}$$

$$= 0.00847$$

(3) n = 100 , p = 0.03

E(x) = np = 100(0.03) = 3

Exercises for Section 2.3
(Exam Questions)

(2.11) A multiple choice test has 5 questions and each question has 4 choices, one of which is correct. You do not have a clue about the subject matter and guess your way through the test. What is the probability you guess**:**

(1) At least one question correctly.

(2) Less than 2 questions correctly.

(3) Exactly 3 questions correctly.

(4) Use **MINITAB** to answer The **Parts (1)**, **(2)** and **(3)**.

(Exam 1999)

(2.12) In a large batch of finished units there are known to be **10%** defectives. If a random sample of **five** units is selected from the batch, calculate the probabilities that the sample will contain**:**

(a) No defectives **(b) One** defective

(c) At most **4** good units

(Exam 2002)

(2.13) **Six** children are born in a maternity hospital on one day. If the probability of having a girl is **0.48**, calculate the following probabilities**:**

(1) The number of **girls** is **3**.

(2) The number of boys is the same as the number of girls.

(3) There will be more boys than girls.

(2.14) A fair die is cast; then n fair coins are tossed, where n is the number shown on the die. What is the probability of exact two heads?

(Exam 2006)

(2.15) From past experience, it is known that 60% of applicants pass on assessment test. In a group of 5 applicants, find the probability that:

(1) All applicants pass the test.

(2) At least one applicant fails the test.

(Exam 2007)

(2.16) Suppose you are given a **three-question** multiple - choice quiz in which each question has **four** optional answers, **one** of which is **correct**.

(1) What is the probability of getting a perfect score if you are forced to guess at each question**?**

Hint: Use two different methods.

(2) Suppose it takes at **least two correct** answers out of **three** to pass the test. What is the probability of **failing** if you are forced to guess at each question**?**

(3) Suppose through some late-night studying you are able to correctly eliminate **two** answers on each question. Now answer **Part (2)**.

(Exam 2008)

(2.17) Calculate the probability that in a group of **5** people

(1) None has his or her **birthday** on a **Saturday**.

(2) Two or more have their **birthdays** on **Friday**.

(Exam 2013)

(2.18) If **x** is a **Binomial** random variable with **n = 4** and **P(x = 2) = 0.3456**, find the value of **p**.

(Exam 2014)

(2.19) A hundred years ago the occupational disease in an industry was such that the workmen had **20%** chance of suffering from it.

(1) If **five** workmen were selected at random, what is the probability that **two** of them will suffer from the disease?

(2) **How many** workmen could have been selected at random before the **probability** that **at least one** of them will suffer from the disease became **greater than 0.9**?

(Exam 2017)

(2.20) A fair die has its faces numbered **1**, **1**, **2**, **2**, **2** and **3**. In a game, this die is cast; then **n** fair coins are tossed where **n** is the number shown on the die. What is the probability of getting exactly **two** heads?

(Exam 2018)

(2.21) Suppose you are taking a **multiple-choice** test with **4** choices for each question, one of which is correct. In answering a question on this test, the **probability** that you **know** the answer is **0.6**. If you **don't know** the answer, you choose one **at random**. What is the **probability** that you **knew the answer** to a question, **given** that you **answered** it **correctly**?

(A) 0.82 **(B) 0.84** **(C) 0.88** **(D) 0.86**

(MCQ Exam 2019)

2.4 Poisson Distribution:

Poisson distribution is actually another discrete probability distribution. The Poisson distribution, like the binomial, is a counted number of times something happens. The difference is that there is no specified number n of possible tries. It describes the probability to find exactly x events in a given length of time if the events occur independently at a constant rate. In addition, the Poisson distribution can be obtained as an approximation of a Binomial distribution when the number of trials n of the latter distribution is large, success probability p is small, and np is a finite number. The average number of successes will be given in a certain interval (this interval could be time, length, area or volume) and is called “Lambda” and denoted by the symbol “**λ**”.

The Poisson distribution was developed by the French mathematician Simeon Denis Poisson in 1837. in statistics, a distribution function useful for characterizing events with very low probabilities of occurrence within some definite time or space.

Characteristics of Poisson Distribution:

The main characteristics which describe Poisson distributions are**:**

1. The experiment consists of counting the number of events that will occur during a specific interval of time or in a specific distance, area, or volume.
2. The probability that an event occurs in a given time, distance, area, or volume is the same.
3. Each event is independent of all other events. For example, the number of people who arrive in the first hour is independent of the number who arrive in any other hour.

Some Applications of Poisson Distributions:

1. The hourly number of customers arriving at a bank.
2. The daily number of accidents on a particular highway.
3. Rare diseases.
4. The number of emergency phone calls received at a center in an hour.
5. Number of typing errors on a page.
6. Failure of a machine in one month.

The Distribution Formula:

Below is the Poisson distribution formula, where the mean (average) number of events within a specific time frame is designed by λ . The probability formula is:

$$\mathbf{P(x) = \frac{e^{-x}\,\lambda^{x}}{x!}}$$

Where **x = 0, 1, 2, ….**

e = 2.71828 (but use your calculator's e button)

λ = Mean number of events per interval

If the variable x follows the with one parameter λ , we write

$$\mathbf{X \sim Po\,(\lambda)}$$

When the Poisson Distribution is Valid?

The Poisson Distribution is only a valid probability analysis tool under certain conditions. It is a valid statistical model if all the following conditions exist:

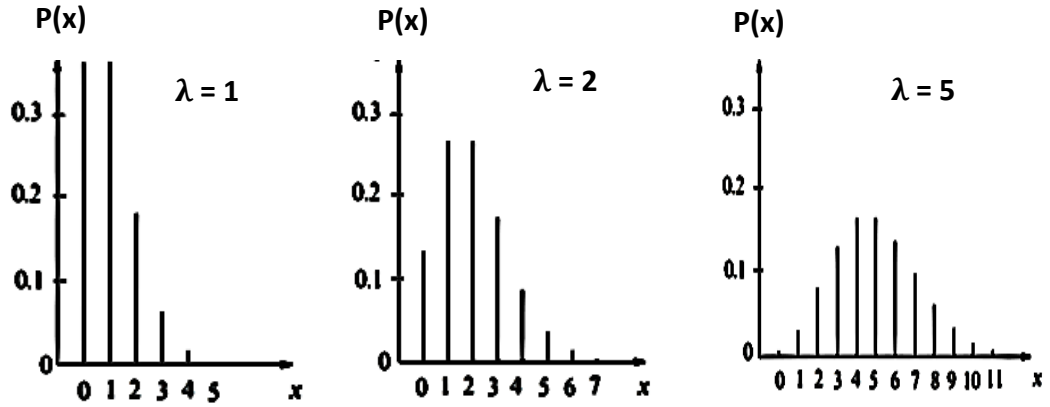
1. x is the number of times an event happens within a specified time period, and the possible values for x are simple numbers such as 0, 1, 2, 3, 4, 5, etc.

2. No occurrence of the event being analyzed affects the probability of the event re-occurring (events occur independently).

3. The event in question cannot occur twice at exactly the same time. There must be some interval of time even if just half a second that separates occurrences of the event.
4. The probability of an event happening within a portion of the total time frame being examined is proportional to the length of that smaller portion of the time frame.
5. The number of trials (chances for the event to occur) is sufficiently greater than the number of times the event does actually occur (in other words, the Poisson distribution is only designed to be applied to events that occur relatively rarely).

Given the above conditions, then x is a random variable, and the distribution of x is a Poisson distribution.

The Shape of the Poisson Distribution



We observe that the Poisson distributions

1. are unimodal
2. exhibit positive skew (that decreases as **λ** increases).
3. are centered roughly on **λ**.
4. have variance (spread) that increases as **λ** increases.

Mean and Variance of Poisson Distribution:

If λ is the average number of successes occurring in a given time interval or region in the Poisson distribution, then the mean and the variance of the Poisson distribution are both equal to λ.

$$\mathbf{E(X) = u = \lambda}$$

and

$$\mathbf{Var\ (x) = \sigma^2 = \lambda}$$

Examples of Poisson Distribution:

Note: **(1)** In a Poisson distribution, only **one** parameter, λ is needed to determine the probability of an event.

(2) It is necessary to use the calculator's button **(e^x)** to perform the calculations for this distribution.

Example (2.17):

A life insurance salesman sells on the average **3** life insurance policies per week. Use Poisson Distribution to calculate the probability that in a given week he will sell:

(1) Some policies

(2) 2 or more policies but **less than 5** policies.

(3) Assuming that there are **5** working days per week, what is the probability that in a given **day** he will sell **one** policy?

Solution:

Here $\lambda = 3$ per week

(1) Some policies mean "1 or more". We can work this out by finding 1 minus the "zero policies" probability.

$P(x > 0) = 1 - P(0)$

$$= 1 - \frac{e^{-3}\ 3^0}{0!} = 1 - 0.0498 = 0.9502$$

(2) The probability of selling 2 or more, but less than 5 policies

$$= P(2 \leq x \leq 5) = P(2) + P(3) + P(4)$$

$$= \frac{e^{-3}\,3^2}{2!} + \frac{e^{-3}\,3^3}{3!} + \frac{e^{-3}\,3^4}{4!}$$

$$= 0.224 + 0.224 + 0.168 = 0.616$$

(3) Average number of policies per day = $\frac{3}{5} = 0.6$

So, on a given day, P(1) = $\frac{e^{-0.6}\,0.6^1}{1!} = 0.3293$

Example (2.18):

Vehicles pass through a junction on a busy road at an average rate of **300** per **hour**.

(1) Find the probability that none passes in a given minute.

(2) What is the expected number passing in two minutes?

(3) Find the probability that this expected number actually pass through in a given two-minute period.

Solution:

(1) The average number of cars per minute ($\boldsymbol{\lambda}$) = $\frac{300}{60} = 5$

Here x = 0 , P(0) = $\frac{e^{-5}\,5^0}{0!} = 0.0067$

(2) The value of $\boldsymbol{\lambda}$ for a two-minute period = 2 × 5 = 10

Expected number passing in 2 minutes = $\lambda = 10$

(3) Now, with $\lambda = 10$, we have

P(10) = $\frac{e^{-10}\,10^{10}}{10!} = 0.1251$

Example (2.19):

If electricity power failures occur according to a Poisson distribution with an average of **3** failures every **twenty weeks.**

(1) Calculate the probability that there will not be more than one failure during a particular **week**.

(2) Find the probability that this expected number actually pass through in a given **two-minute** period.

Solution:

(1) The average number of failures per week is $\frac{3}{20}$ = 0.15

"Not more than one failure" means we need to include the probabilities for "0 failure" plus "1 failure".

$$P(0) + P(1) = \frac{e^{-0.15}\ 0.15^{0}}{0!} + \frac{e^{-0.15}\ 0.15^{1}}{1!}$$

$$= 0.8607 + 0.1291 = 0.9898$$

Example (2.20):

As only **3** students came to attend the class today, find the probability for exactly 4 students to attend the classes tomorrow.

Solution:

Given**:** Average rate per day (λ) = 3 and x = 4

$$P(4) = \frac{e^{-3}\ 3^{4}}{4!} = 0.168$$

Example (2.21):

Suppose we know that births in a hospital occur randomly at an average rate of 1.8 births per hour. What is the probability that we observe 5 births in a given two - hour interval?

Solution:

If births occur randomly at a rate of 1.8 births per 1 hour interval, then births occur randomly at a rate of 3.6 births per two - hour interval.

Thus, λ = 2 × 1.8 = 3.6 and x = 5

$$P(5) = \frac{e^{-3.6}\ 3.6^5}{5!} = 0.1377$$

Example (2.22):

A customer service center receives about ten emails every half - hour. What is the probability that the customer service center receives more than one email in the next six minutes?

Solution:

The average number of emails per minute = $\frac{10}{30} = \frac{1}{3}$

Thus, the value of λ for six-minute interval = $6\left(\frac{1}{3}\right) = 2$

$$P(x > 1) = 1 - \{P(0) + P(1)\} = 1 - \left\{\frac{e^{-2}\ 2^0}{0!} + \frac{e^{-2}\ 2^1}{1!}\right\}$$

$$= 1 - \{0.1353 + 0.2707\} = 0.594$$

Example (2.23):

The number of industrial injuries per working week in a particular factory is known to follow a Poisson distribution with **mean 0.5**. Find the probability that:

(a) In a particular week there will be:

(i) Less than **2** accidents,

(ii) More than **2** accidents;

(b) In a three-week period there will be no accidents.

Solution:

Let x be 'the number of accidents in one week, So **X ~ Po (0.5)**

(a) (i) $P(x < 2) = P(0) + P(1)$

$$= \frac{e^{-0.5}\ 0.5^0}{0!} + \frac{e^{-0.5}\ 0.5^1}{1!}$$

$$= 0.6065 + 0.3033 = 0.9098$$

(ii) P(x > 2) = 1 - [P(0) + P(1) + P(2)]

$$= 1 - \left[\frac{e^{-0.5}\, 0.5^{0}}{0!} + \frac{e^{-0.5}\, 0.5^{1}}{1!} + \frac{e^{-0.5}\, 0.5^{2}}{2!} \right]$$

$$= 1 - [\, 0.6065 + 0.3033 + 0.0758]$$

$$= 1 - 0.9856 = 0.0144$$

(b) The average number of accidents for three - week period

$$= 3 \times 0.5 = 1.5$$

$$P(0) = \frac{e^{-1.5}\, 1.5^{0}}{0!} = 0.2231$$

Example (2.24):

Patients arrive at a hospital accident and emergency department at random at a rate of **6** per hour.

(a) Find the probability that, during any **90 - minute**s period, the number of patients arriving at the hospital accident and emergency department is

(i) exactly 2 **(ii)** at most 2

(b) A patient arrives at 11.30 a.m., find the probability that the next patient arrives before 11.45 a.m.

Solution:

The average number of patients arriving **per minute** = $\frac{6}{60} = 0.1$

(a) The value of λ for a 90 - minute period = $90 \times 0.1 = 9$

(i) $P(2) = \frac{e^{-9}\, 9^{2}}{2!} = 0.005$

(ii) P(at most 2) = $P(x \leq 2) = P(0) + P(1) + P(2)$

$$= \frac{e^{-9}\, 9^{0}}{0!} + \frac{e^{-9}\, 9^{1}}{1!} + \frac{e^{-9}\, 9^{2}}{2!}$$

= 0.00012 + 0.00111 + 0.005 = 0.0062

(b) The value of **λ** for a 15 - minute period (11.45 a.m. – 11.30 a.m.) = 15 × 0.1 = 1.5

Here x = 1, P(1) = $\frac{e^{-1.5}\ 1.5^{1}}{1!}$ = 0.3347

Exercises for Section 2.4
(Exam Questions)

(2.22) Serious accidents occur at random in a particular manufacturing industry at the rate of 1.5 per week.

(a) Find the probability that exactly one accident will occur in a week.

(b) Find the probability of less than two accidents occurring during a four-week period.

(c) Use **MINITAB** to answer **Parts (a)** and **(b)**.

(Exam 2000)

(2.23) The number of emergency admissions each day to a hospital is found to have a Poisson distribution with mean **2**.

(a) Find the probability that on a particular day there will be **no** emergency admissions.

(b) At the beginning of one day the hospital has **two** beds available for emergencies. Calculate the probability that this will be an insufficient number for the day.

(c) Calculate the probability that there will be exactly **three** admissions altogether on **two** consecutive days.

(Exam 2003 Qena)

(2.24) The average number of faults in a meter of cloth produced by a particular machine is **0.1**.

(a) Find the following probabilities**:**

(i) a length of **4** meters is **free** from faults.

(ii) a length of **one** meter has only **two** faults.

(Exam 2003 Sohag)↓

(b) How long would a piece have to be before the probability that it contains **no** flaws is less than **0.95?**

(Exam 2003 Sohag)

(2.26) Phone calls arrive at the rate of 48 per hour at the reservation desk for Regional Airways.

(1) Compute the probability of receiving exactly 10 calls in 15 minutes.

(2) Find the probability of receiving 3 calls in a five-minute interval of time.

(3) Suppose no calls are currently on hold. If the agent takes 5 minutes to complete the current call, how many calls do you expect to be waiting for that time? What is the probability that none will be waiting?

(Exam 2007)

(2.28) Suppose the **average** number of accidents occurring **weekly** on a particular highway is equal to **1.2**. Find the following probabilities:

(1) Exactly **one** accident will occur in a **week.**

(2) **At least one** accident occurring during a **four-week** period.

(Exam 2014)

(2.29) Based on past experience, it is assumed that the number of flaws per foot in rolls of grade 2 paper follows a **Poisson** distribution with an **average** of **1** flaw per **5** feet of paper. What is the probability that in a ……

(1) **1-foot** roll there will be **2** flaws.

(2) **10-feet** roll there will be **at least 1** flaw.

(Exam 2015)

(2.30) A computer center manager reports that his computer system experienced **three** component **failures** during the past **100** days.

(1) Find the probability of **no failures** in a given **day**.

(2) What is the probability of **at least two** failures in a **three-day** period? **(Exam 2016)**

(2.31) Suppose the number of demands for taxis to the Uber firm is **Poisson** distributed with the **mean** of **four** demands in **30** minutes. Find the probabilities of**:**

(1) **No calls** in **30** minutes **(2)** **Two** calls in **1 hour**.

(Exam 2017)

(2.32) **Car arriving** for gasoline at a Shell service station follow a **Poisson** distribution with a **mean** of **3** per **hour**.

(1) Determine the **probability** that over the next **hour** only **one** car will arrive.

(2) Find the **probability** that in the next **2** hours **more than one** car will arrive.

(3) What is the **time** required so that the **probability** of **not** arriving **any** car is **0.6**? **(Exam 2018)**

(2.33) The **average** number of **faults** in a **meter** of cloth produced by a particular machine is **0.2**. **(MCQ)**

(1) The probability that a length of **4** meters is **free** from faults is

(A) **0.45** **(B)** **0.52** **(C)** **0.48** **(D)** **0.56**

(2) **How long** would a **piece** have to be such that the **probability** that it contains **no** faults is **0.0183**?

(A) **18 m** **(B)** **19 m** **(C)** **20 m** **(D)** **22 m**

(MCQ Exam 2019)

2.5 Hypergeometric Distribution:

In Section 2.2, we presented Binomial experiments. Recall, the Binomial probability distribution can be used to compute the probabilities of experiments when there are a fixed number of trials in which there are two mutually exclusive outcomes and the probability of success for any trial is constant. In addition, the trials must be independent. When small samples are obtained from large finite populations, it is reasonable to assume independence of events. That is, when obtaining a sample of size n from a population whose size is N, we are willing to assume independence of the events provided that (the sample size is less than 5% of the population size).

What if the requirement of independence is not satisfied? Under these circumstances, the experiment is a hypergeometric experiment.

Hypergeometric distribution, in statistics, distribution function in which selections are made from two groups without replacing members of the groups. The hypergeometric distribution differs from the binomial distribution in the lack of replacements. A simple everyday example would be the random selection of members for a team from a population of girls and boys.

In symbols, let the size of the population selected from be N, with S elements of the population belonging to one group (for convenience, called successes) and N − S belonging to the other group (called failures). Further, let the number of samples drawn from the population be n, such that $0 \leq n \leq N$. Then the probability (P) that the number (x) of elements drawn from the successful group is equal to some number (x) is given by the following Hypergeometric distribution formula**:**

$$P(X = x) = P(x) = \frac{\binom{S}{x}\binom{N-S}{n-x}}{\binom{N}{n}}, \quad 0 \le x \le S, \quad 0 \le n - x \le N - S$$

Where

N = the size of the population

n = the sample size or the number of trials (draws)

S = the number of successes in the population

N-S = the number of non-successes in the population

X = the number of successes of interest. It may be 0, 1, 2, 3,...

The logic for the Hypergeometric distribution was developed in **Section 2.3** using the classic definition of probability and the counting formulas for combinations. In the above formula of the Hypergeometric distribution the individual components are:

1. The number of possible ways that x successes can be selected for the sample out of S successes contained in the population is:

$$\binom{S}{x} = \frac{S!}{x!(S-x)!}$$

2. The number of possible ways that **n - x** non-successes can be selected from the population that contains **N - S** non-successes.

$$\binom{N-S}{n-x} = \frac{(N-S)!}{(n-x)!(N-S-n+x)!}$$

3. Finally the total number of different samples of size **n** that can be obtained from a population of size N is:

$$\binom{N}{n} = \frac{N!}{n!(N-n)!}$$

In general, if the random variable x follows a Hypergeometric distribution with the **3** parameters **N**, **s** and **n**, we write

$$X \sim H(N, S, n)$$

The Mean (Expected Value) and Variance of the Hypergeometric Distribution:

The expected value (mean) and variance of the Hypergeometric distribution are given by:

$$E(X) = \mu = \left(\frac{nS}{N}\right),$$

$$Var(X) = \sigma^2 = n\left(\frac{S}{n}\right)\left(\frac{N-S}{N}\right)\left(\frac{N-n}{N-1}\right)$$

Criteria for a Hypergeometric Experiment:

A probability experiment is said to be a hypergeometric experiment provided**:**

1. The finite population to be sampled has N elements.

2. For each trial of the experiment, there are two possible outcomes, success or failure. There are exactly s successes in the population.

3. A sample of size n is obtained from the population of size N without replacement. That is, The probability of success changes after each trial.

If a probability experiment satisfies these three requirements, the random variable X, the number of successes in n trials of the experiment, follows the Hypergeometric probability distribution.

Note:

(1) In a Hypergeometric distribution, **three parameters**, **N**, **S**, and **n** are needed to determine the probability of an event.

(2) It is necessary to use the calculator's button **(^nC_r)** to perform the calculations for this distribution.

Example (2.25):

The classical application of the hypergeometric distribution is **sampling without replacement**. Think of an urn with two

colors of marbles, red and green. Define drawing a green marble as a success and drawing a red marble as a failure (analogous to the Binomial distribution). If N describes the number of **all marbles in the urn** (see contingency table below) and S describes the number of **green marbles**, then N – S corresponds to the number of **red marbles**. In this example, X is the random variable whose outcome is x, the number of green marbles actually drawn in the experiment. This situation is illustrated by the following contingency table**:**

Color	Drawn	Not Drawn	Total
Green marbles	x	S − x	S
Red marbles	n − x	N + x − n − S	N − S
Total	n	N − n	N

Now, assume (for example) that there are 5 green and 45 red marbles in the urn. Standing next to the urn, you close your eyes and draw 10 marbles without replacement. What is the probability that exactly 4 of the 10 are green? Note that although we are looking at success/failure, the data are not accurately modeled by the binomial distribution, because the probability of success on each trial is not the same, as the size of the remaining population changes as we remove each marble.

This problem is summarized by the following contingency table**:**

Color	Drawn	Not Drawn	Total
Green marbles	x = 4	S − x = 1	S = 5
Red marbles	n − x = 6	N + x − n – S = 39	N − S = 45
Total	n = 10	N − n = 40	N = 50

The probability of drawing exactly *x* green marbles can be calculated by the formula

$$P(x) = \frac{\binom{S}{x}\binom{N-S}{n-x}}{\binom{N}{n}}$$

Hence, in this example calculate

$$P(4) = \frac{\binom{5}{4}\binom{45}{6}}{\binom{50}{10}} = 0.004$$

Intuitively we would expect it to be even more unlikely that all 5 green marbles will be among the 10 drawn.

$$P(5) = \frac{\binom{5}{5}\binom{45}{5}}{\binom{50}{10}} = 0.0001$$

As expected, the probability of drawing 5 green marbles is roughly 35 times less likely than that of drawing 4.

Example (2.26):

A deck of cards contains 20 cards; **6 red** cards and **14 black** cards. **5** cards are drawn randomly without replacement. What is the probability that exactly **4 red** cards are drawn?

Solution:

In this example: **N = 20 , n = 5 , S = 6 , x = 4**

The probability of choosing exactly 4 red cards is:

$$P(4) = \frac{\binom{6}{4}\binom{14}{1}}{\binom{20}{5}} = 0.0135$$

Where

- $\binom{6}{4}$ means that out of 6 possible red cards, we are choosing 4.

- $\binom{14}{1}$ means that out of a possible 14 black cards, we are choosing1.
- $\binom{20}{5}$ means that out of 20 cards, we are choosing 5 cards.

The Binomial distribution doesn't apply here, because the cards are not replaced once they are drawn. In other words, the trials are not independent events. For example, for one red card, the probability is 6/20 on the first draw. If that card is red, the probability of choosing another red card falls to 5/19.

Example (2.27):

A small voting district has 100 female voters and 50 male voters. A random sample of 10 voters is drawn. What is the probability that exactly 7 of the voters will be female?

Solution:

Here **N = 196 , n = 10 , S = 101 , x = 7**

$$P(7) = \frac{\binom{100}{7}\binom{50}{3}}{\binom{150}{10}} = 0.2683$$

Example (2.28): Quality Control

A container has 100 items 5 of which the worker who packed the container knows are defective. A merchant wants to buy the container without knowing the above information. However, he will randomly pick 20 items from the container and will accept the container as good if there is at most one bad item in the selected sample. What is the probability that the merchant will declare the container to be good?

Solution:

In this Example**: N = 100 , n = 20 , S = 5 , x = 0 or 1**

Let **A** be the event that there is no defective item in the selected sample, and

B the event that there is exactly one defective item in the selected sample.

Since A and B are mutually exclusive events we obtain the following results:

P(The merchant will declare the container to be good) is equal to

$$P(A) + P(B) = \frac{\binom{95}{20}\binom{5}{0}}{\binom{100}{20}} + \frac{\binom{95}{19}\binom{5}{1}}{\binom{100}{20}} = 0.3193 + 0.4201$$

$$= 0.7394$$

Example (2.29):

We repeat **Example (2.28)** with a different sample size. Instead of testing the quality of the container with a sample size of 20 the merchant decides to test it with a sample size of 50. As before, he will accept the container as good if there is at most one bad item in the selected sample. What is the probability that the merchant will declare the container to be good?

Solution:

In this case: **N = 100 , n = 50 , S = 5 , x = 0 or 1**

P(The merchant will declare the container to be good) is equal to

$$P(A) + P(B) = \frac{\binom{95}{50}\binom{5}{0}}{\binom{100}{50}} + \frac{\binom{95}{49}\binom{5}{1}}{\binom{100}{50}} = 0.0281 + 0.1529 = 0.181$$

The above examples illustrate the impact of the sample size on decisions that can be made. With a bigger sample size, the merchant is more likely to make a better decision because there is a greater probability that more of the defective items will be included in the sample. For example, if the decision was based on the result of a random sample of **10** items, we would get **P(A) +P(B) = 0.9231**. Similarly, if it was based on the result of a random sample of **40** items, we would get **P(A) + P(B) = 0.3316**, while with a sample of **60** items, we would get **P(A) + P(B) = 0.0816**. Note that while a bigger sample size has the tendency to reveal the fact that the container has a few bad items, it also involves more testing. Thus, to get better information we must be prepared to do more testing, which is a basic rule in quality control.

Example (2.30):

A certain library has a collection of 10 books on probability theory. Six of these books were written by American authors and four were written by British authors.

a. If you randomly select one of these books, what is the probability that it was written by an American author?

b. If you select five of these books at random without replacement, what is the probability that two of them were written by American authors and three of them were written by British authors?

Solution:

a. There are $\binom{6}{1}$ = 6 ways to choose a book written by an American author and $\binom{10}{1}$ = 10 ways to choose a book at random. Therefore, the probability that a book chosen at random was written by an American author is:
p = 6/10 = 0.6.

b. This is an example of the hypergeometric distribution in which **N = 10 , n = 5 , S = 6 , x = 2**

Thus, the probability that of the five of these books selected at random, two of them were written by American authors and three of them were written by British authors is given by

$$P(2) = \frac{\binom{6}{2}\binom{4}{3}}{\binom{10}{5}} = 0.2381$$

Example (2.31):

Suppose we randomly select 5 cards without replacement from an ordinary deck of playing cards. What is the probability of getting exactly 2 red cards (i.e., hearts or diamonds)?

Solution:

This is a hypergeometric experiment in which we know the following**:**

N = 52; since there are 52 cards in a deck.

n = 5; since we randomly select 5 cards from the deck.

S = 26; since there are 26 red cards in a deck.

x = 2; since 2 of the cards we select are red.

Since $$P(x) = \frac{\binom{S}{x}\binom{N-S}{n-x}}{\binom{N}{n}}$$

We plug these values into the hypergeometric formula as follows:

$$P(2) = \frac{\binom{26}{2}\binom{26}{3}}{\binom{52}{5}} = 0.3251$$

Thus, the probability of randomly selecting 2 red cards is 0.3251.

Example (2.32):

Suppose we want to find the probability that a committee of 10 people chosen from a group consisting of 40 teachers, and 20 students, will include six teachers (four students).

Solution:

Here, **N = 60 , n = 10 , S = 40 , x = 5**

Since $$P(x) = \frac{\binom{S}{x}\binom{N-S}{n-x}}{\binom{N}{n}}$$

Substituting these values into this formula, we get

$$P(2) = \frac{\binom{40}{6}\binom{20}{4}}{\binom{60}{10}} = 0.2467$$

Thus, the probability that the committee will include 6 teachers is 0.2467.

Example (2.33):

Suppose a shipment of 100 DVD players is known to have ten defective players. An inspector randomly chooses 12 for inspection. He is interested in determining the probability that, among the 12 players, at most two are defective.

Solution:

The two groups are the 90 non-defective DVD players and the 10 defective DVD players. The group of interest (first group) is the defective group because the probability question asks for the probability of at most two defective DVD players. The size of the sample is 12 DVD players. (They may be non-defective or defective.) Let X = the number of defective DVD players in the

sample of 12. X takes on the values 0, 1, 2 ,…,10, and S = 0, 1, 2,…,10. X may not take on the values 11 or 12. The sample size is 12, but there are only 10 defective DVD players.

Writing the probability statement mathematically**:**

Here N = 100 , n = 12 , S = 10 , x = 0, 1, or 2

Plugging these values into the formula of the Hypergeometric distribution, we get

Since $$P(x) = \frac{\binom{S}{x}\binom{N-S}{n-x}}{\binom{N}{n}}$$

$$P(x \leq 2) = P(0) + P(1) + P(2)$$

$$= \frac{\binom{10}{0}\binom{90}{12}}{\binom{100}{12}} + \frac{\binom{10}{1}\binom{90}{11}}{\binom{100}{12}} + \frac{\binom{10}{2}\binom{90}{10}}{\binom{100}{12}}$$

$$= 0.2608 + 0.3961 + 0.2451 = 0.9092$$

Example (2.34):

A palette has 200 milk cartons. Of the 200 cartons, it is known that ten of them have leaked and cannot be sold. A stock clerk randomly chooses 18 for inspection. He wants to know the probability that among the 18, no more than two are leaking. Give four reasons why this is a hypergeometric problem. Then, find the required probability.

Solution:

The four reasons are:

- There are two groups.
- You are concerned with a group of interest.
- You sample without replacement.
- Each pick is not independent.

For obtaining the required probability $P(x \leq 2)$:

$N = 200$, $n = 18$, $S = 10$, $x = 0$, 1, or 2

$P(x \leq 2) = P(0) + P(1) + P(2)$

$$= \frac{\binom{10}{0}\binom{190}{18}}{\binom{200}{18}} + \frac{\binom{10}{1}\binom{190}{17}}{\binom{200}{18}} + \frac{\binom{10}{2}\binom{190}{16}}{\binom{200}{18}}$$

$= 0.3806 + 0.3960 + 0.1741$

$= 0.9507$

Exercises for Section 2.5

(2.34) A manufacturer received an order of 250 computer chips. Unfortunately, 12 of the chips are defective. To test the shipment, the quality-control engineer randomly selects 20 chips from the box of 250 and tests them. The random variable X represents the number of defective chips in the sample.

(a) What is the probability of obtaining 4 defective chips?

(b) What is the probability of obtaining 3 defective chips?

(c) What is the probability that the quality-control engineer will not find any defective chips?

(d) What is the probability of obtaining 14 defective chips?

(e) How many defective chips would you expect to select?

(2.35) Baseball Lineup A baseball team has 25 players, 7 of whom bat left-handed. Suppose that the manager of this team is frustrated with the way the team is playing, so he decides to randomly select 9 players to play in the upcoming game. The random variable X will be the number of left-handed batters in the game.

(a) What is the probability of creating a lineup with 2 lefties?

(b) What is the probability of creating a lineup with 1 lefty?

(c) What is the probability of creating a lineup with no lefties?

(d) What is the probability of creating a lineup with 8 lefties?

(e) How many lefties would you expect to find in the lineup?

(2.36) In a small pond there are 50 fish, 10 of which have been tagged. A fisherman's catch consists of 7 fish (assume his catch is a random selection done without replacement). What is the probability that exactly 2 tagged fish are caught?

(2.37) A bag contains 24 balls, of which 12 are red and 12 are black. Let X be the number of red balls in a sample of 5 balls selected at random and without replacement from the bag. Find:
(a) $P(x = 2)$ **(b)** $P(x \leq 2)$ **(C)** $\mu = E(X)$ **(d)** $\sigma^2 = Var(X)$

(2.38) A bag contains letter tiles. Forty-four of the tiles are vowels, and 56 are consonants. Seven tiles are picked at random. You want to know the probability that four of the seven tiles are vowels. What is the group of interest, the size of the group of interest, and the size of the sample?

(2.39) A gross of eggs contains 144 eggs. A particular gross is known to have 12 cracked eggs. An inspector randomly chooses 15 for inspection. What is the probability that, among the 15, at most three are cracked? What is X, and what values does it take on?

(2.40) You are president of an on-campus special events organization. You need a committee of seven students to plan a special birthday party for the president of the college. Your organization consists of 18 women and 15 men. You are interested in the number of men on your committee. If the members of the committee are randomly selected, what is the probability that your committee has more than four men?

(2.41) For Exercises **(2.34) to (2.36)**, compute the mean and standard deviation of the hypergeometric random variable X.

(2.42) A school site committee is to be chosen randomly from six men and five women. If the committee consists of four members chosen randomly, what is the probability that two of them are men? How many men do you expect to be on the committee?

(2.43) A basketball team is to be chosen randomly from 15 boys and 12 girls. Ten players were randomly selected. What is the probability that eight of the players will be boys? What is the group of interest and the sample?

Chapter (3)
Continuous Probability Distributions

Contents

Chapter 3
Continuous Random Variables

3.1 Continuous Variable:

Continuous Variable:
A variable that can assume any numerical value over a certain interval.

The time taken to complete an examination is an example of a continuous variable because it can assume any value, let us say, between 30 and 60 minutes. The time taken may be 42.6 minutes, 42.67 minutes, or 42,674 minutes. (Theoretically, we can measure time as precisely as we want). Similarly, the age of a person. Neither time nor age can be counted in a discrete fashion. That is, a **continuous random variable** which is a random variable where the data can take infinitely many values.

Examples:
- Height of students in class
- Weight of students in class
- Time it takes to get to school
- Age of persons
- Amount of money

If a random variable is a continuous variable, its probability distribution is called a **continuous probability distribution**.
A continuous probability distribution differs from a discrete probability distribution in several ways:

- The probability that a continuous random variable will assume a particular value is zero.
- As a result, a continuous probability distribution cannot be expressed in tabular form.
- Instead, an equation or formula is used to describe a continuous probability distribution.

Probability Distribution of Continuous Random Variables:

For a discrete random variable x the probability that x assumes one of its possible values on a single trial of the experiment makes good sense. This is not the case for a continuous random variable. For example, suppose x denotes the length of time a commuter just arriving at a bus stop has to wait for the next bus. If buses run every 30 minutes without fail, then the set of possible values of x is the interval denoted [0,30], the set of all decimal numbers between 0 and 30. But although the number 7.211916 is a possible value of x, there is little or no meaning to the concept of the probability that the commuter will wait precisely 7.211916 minutes for the next bus. If anything, the probability should be zero, since if we could meaningfully measure the waiting time to the nearest millionth of a minute it is practically inconceivable that we would ever get exactly 7.211916 minutes. More meaningful questions are those of the form: What is the probability that the commuter's waiting time is less than 10 minutes, or is between 5 and 10 minutes? In other words, with continuous random variables one is concerned not with the event that the variable assumes a single particular value, but with the event that the random variable assumes a value in a particular interval.

Definition:

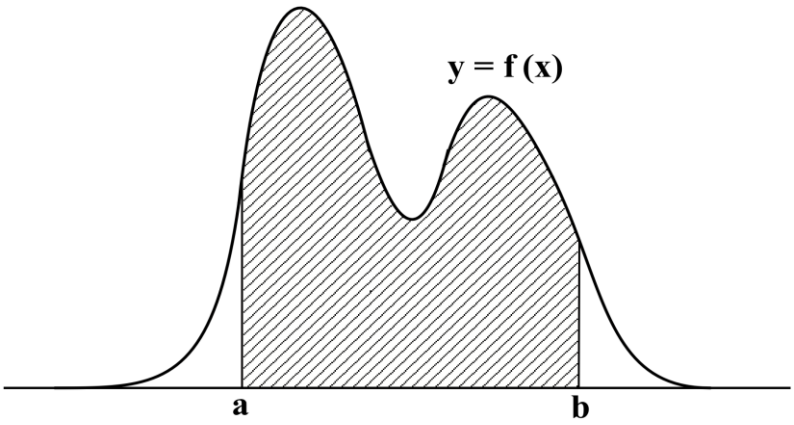
Probability Density Function:

The equation used to describe a continuous probability distribution is called a probability density function. Sometimes, it is referred to as a density function, a PDF, or a pdf.

The probability that X assumes a value in the interval [a,b] is equal to the area of the region that is bounded above by the graph of the equation y = f(x), bounded below by the x-axis, and bounded on the left and right by the vertical lines through **a** and **b**, as

illustrated in the following figure "Probability Given as Area of a Region under a Curve".

"Probability Given as Area of a Region under a Curve"
P(a < X < b) = Area of Shaded Region



The density function of a continuous random variable has the following properties:

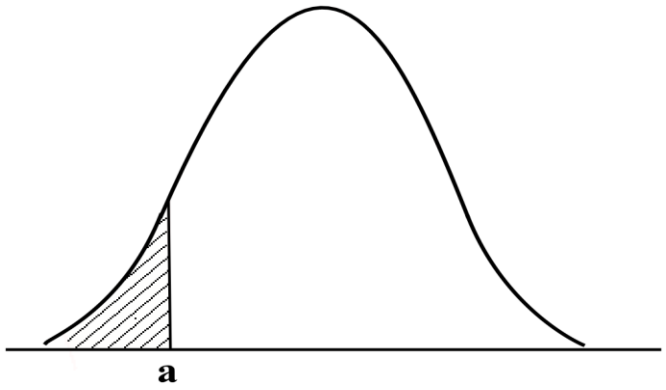
- Since the continuous random variable is defined over a continuous range of values (called the domain of the variable), the graph of the density function will also be continuous over that range.
- The area bounded by the curve of the density function and the x - axis is equal to 1, when computed over the domain of the variable, that is

$$\int_{\text{all } x} \mathbf{f(x)dx = 1}$$

- The probability that a random variable assumes a value between **a** and **b** is equal to the area under the density function bounded by **a** and **b**.

Now, consider the probability density function shown in the graph below. Suppose we wanted to know the probability that the random variable X was less than or equal to a. The probability that X is less than or equal to *a* is equal to the area under the

curve bounded by *a* and minus infinity as indicated by the shaded area.



Examples (3.1):

x is a continuous random variable with probability density function given by

$$f(x) = cx \text{ for } \quad 0 \leq x \leq 1,$$

where **c** is a constant. Find the value of **c**.

Solution:

If we integrate f(x) between 0 and 1 we get $\frac{c}{2}$.

Hence $\frac{c}{2} = 1$ (from the useful fact above!), giving c = 2.

Cumulative Distribution Function (c.d.f.):

If X is a continuous random variable with p.d.f. f(x) defined on $a \leq x \leq b$, then the cumulative distribution function (c.d.f.), written F(t) is given by:

$$F(t) = P(x \leq t) = \int_a^t f(x)dx$$

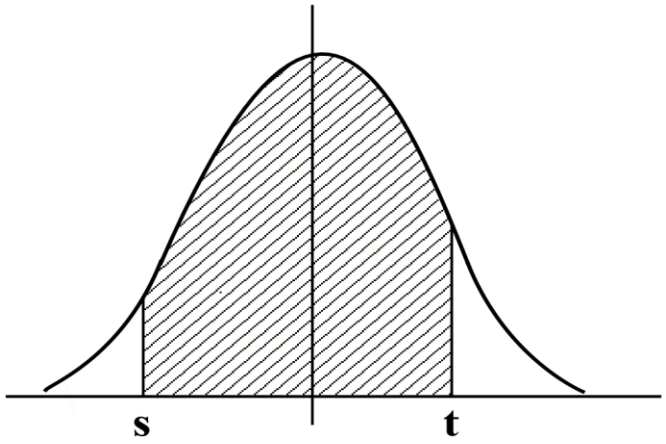
Where **a** is the **minimum** value of the random variable x.

So, the c.d.f. is found by integrating the p.d.f. between the minimum value of X and t.

Similarly, the probability density function of a continuous random variable can be obtained by differentiating the cumulative distribution.

The c.d.f. can be used to find out the probability of a random variable being between two values**:**

P(s ≤ X ≤ t) = the probability that X is between s and t. But this is equal to the probability that X ≤ t minus the probability that X ≤ s. [We want the probability that X is in the shaded area]



Hence:

$$P(s \le X \le t) = P(X \le t) - P(X \le s)$$
$$= \int_s^t f(x)dx - \int_s^t f(x)dx$$
$$= F(t) - F(s)$$

Hint: As mentioned before, for a continuous random variable the probability of any particular value is zero; thus,

$$\mathbf{P(a \le X \le b) = P(a < x < b) = F(b) – F(a)}$$

This result shows that the probability of a random variable assuming value in any interval is the same whether or not the endpoints are included.

Expectation and Variance:

With discrete random variables, we had that the expectation was $\mathbf{\Sigma\{xP(X = x)\}}$, where P(X = x) was the p.d.f. It may come as no

surprise that to find the expectation of a continuous random variable, we integrate rather than sum.

$$\mathbf{E(x) = \int_{all\ x} xf(x)dx}$$

As with discrete random variables,

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

Example (3.2):

Consider the function:

$$f(x) = \begin{cases} 2x & 0 \le x \le b \\ 0 & \text{otherwise} \end{cases}$$

(a) For what value of b is f(x) a valid PDF?
(b) Find the probability: P(x < 0.5).

Solution:

(a) Since $\int_0^b 2xdx = 1$, then

$[X^2]_0^b = b^2 = 1$, which gives b = 1

(b) P(X < 0.5) = $\int_0^{0.5} 2xdx = 0.25$

Example (3.3):

The PDF of a random variable of X is given by:

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

(a) Find P(0.2 < X < 0.8).

(b) Find the probability P(0.6 < X < 1.2).

Solution:

(a) P(0.2 < x < 0.8) = $\int_{0.2}^{0.8} xdx$

$$= \frac{(0.8)^2}{2} - \frac{(0.2)^2}{2} = 0.3$$

P(0.6 < X < 1.2) = F(1.2) – F(0.6)

F(1.2) = $\int_0^1 x dx + \int_1^{1.2} (2 - x) dx$

Integrating X gives $\frac{X^2}{2}$, and

Integrating (2 – X) gives 2X – $\frac{X^2}{2}$ giving

F(1.2) = {2(1) – $\frac{(1)^2}{2}$ } - {$\frac{(1)^2}{2}$ – 0)} + {2(1.2) – $\frac{(1.2)^2}{2}$}

= {(2.4 – 0.72) – (2 – 0.5)} + 0.5

= (1.68 -1.5) + 0.5 = 0.68

F(0.6) = $\frac{(0.6)^2}{2} - \frac{(0)^2}{2}$ = 0.18

Thus, P(0.2 < X < 0.8) = F(1.2) – F(0.6) = 0.68 – 0.18 = 0.5

Example (3.4):

Let x be a continuous random variable with PDF

$$f(x) = \begin{cases} x^2(2x + 1.5) & 0 < x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

If y = 2/x + 3 , find Var(y).

Solution:

First, note that: Var(y) = Var (2/x + 3) = 4Var $\left(\frac{1}{X}\right)$

Thus, it suffices to find Var$\left(\frac{1}{X}\right)$. So,

$$E\left(\frac{1}{X}\right) = \int_0^1 \left(\frac{1}{x}\right)(x^2)(2x + 1.5)dx = \int_0^1 x(2x + 1.5)dx$$

= ($\frac{2X^3}{3}$ + 0.75X^2) from 0 to 1 giving 17/12

$$E\left(\frac{1}{X^2}\right) = \int_0^1 (2x + 1.5)dx = \frac{5}{2}$$

Thus,

$$Var\left(\frac{1}{X}\right) = E\left(\frac{1}{X^2}\right) - \left\{E\left(\frac{1}{X}\right)\right\}^2$$

$$Var\left(\frac{1}{X}\right) = \frac{5}{2} - \left(\frac{17}{12}\right)^2 = \frac{71}{144}$$, So, we obtain

$$Var(y) = 4\left(\frac{71}{144}\right) = \frac{71}{36}$$

Example (3.5):

Consider the following function**:**

$$f(x) = \begin{cases} x & -2 \le x \le 2 \\ 1 & 2 < x < 3 \\ 0 & \text{otherwise} \end{cases}$$

Is it a valid probability density function for a continuous random variable?

Solution:

$$\int_{-2}^{2} x dx + \int_{2}^{3} x dx = \left[\frac{X^2}{2}\right]_{-2}^{2} + [X]_2^3$$

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

Since $\int_{-2}^{3} f(x)dx = 1$ **,** then it is a valid PDF.

Example (3.6):

The density function of a continuous random variable is

$$f(x) = \begin{cases} 4x(1- x^2) & 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

(a) Find the expected value (mean).
(b) Find P(x ≤ 0.5).

Solution:

(a) $E(x) = \int_0^1 x(4x)(1 - x^2)dx = \left[\left(\frac{4X^3}{3}\right) - \left(\frac{4X^5}{5}\right)\right]_0^1$

$= 4\left(\frac{1}{3}\right) - 4\left(\frac{1}{5}\right) = \frac{8}{15} = 0.533$

(b) $P(x \le 0.5) = \int_0^{0.5} (4x)(1 - x^2)dx = \left[\left(\frac{4X^2}{2}\right) - \left(\frac{4X^4}{4}\right)\right]_0^{0,5}$

$= 2(0.5)^2 - (0.5)^4 = 0.4375$

Exercises for Section 3.1

(3.1) The following density function describes the random variable x:

$$f(x) = \begin{cases} \frac{x}{25} & 0 < x < 5 \\ \frac{10 - x}{25} & 5 < x < 10 \end{cases}$$

(a) Find the probability that x lying between 1 and 3**.**
(b) Compute the probability that x less than 7.
(c) Find the probability that X is greater than 3.

(3.2) The following function density function for the random variable X.

$$f(x) = \frac{x - 1}{8}, \quad 1 < x < 5$$

(a) Find the probability that X lies between 2 and 4.
(b) What is the probability that X is less than 3?
(c) Find the mean of X.

(3.3) A random variable has the following density function**:**

$$f(x) = 1 - 0.5x \ , \ 0 < x < 2$$

(a) Verify that f(x) is a density function.
(b) What is the probability that X is less than 0.5?
(c) Find P(X > 1). **(d)** Find P(x = 1.5).

(3.4) Suppose the random variable X has the PDF

$$f(x) = ax^3, \ 0 < x < 1$$

(a) What is the value of a?
(b) What is the expected value of x?

(c) What is the variance of x?

(d) What is the value of m so that P(x ≤ m) = 1/2?

(3.5) The PDF of a random variable X is given by

$$f(x) = \frac{4\left(9 - x^2\right)}{81}\ ,\ 0 \le x \le 3$$

Find the **mean** and **variance** of x.

(3.6) The CDF of a random variable X is

$$f(x) = \begin{cases} 1 - e^{-2x} & x \ge 0 \\ 0 & x < 0 \end{cases}$$

Find

(a) The density function f(x).

(b) The probability that x > 2.

(c) The probability that Find -3 < x ≤ 4.

(3.7) Suppose the random variable X has the PDF:

$$f(x) = \begin{cases} 0.5x - 0.25x & 0 \le x \le 2 \\ 0.5x + 0.25x & -2 \le x \le 0 \\ 0 & \text{otherwise} \end{cases}$$

Find P(-1 ≤ x ≤ 1).

Hint: You can use symmetry to find this probability.

3.2 Normal Distribution

In probability theory, a normal (or Gaussian) distribution is a type of continuous probability distribution for a real-valued random variable. The general form of its probability density function is

$$\mathbf{f(x)} = \frac{\mathbf{1}}{\boldsymbol{\sigma}\sqrt{\mathbf{2\pi}}}\mathbf{e}^{-\frac{(\mathbf{x}-\boldsymbol{\mu})^{\mathbf{2}}}{\mathbf{2\sigma^{2}}}}$$

Where

- **f(x) = Probability distribution**
- **x = Value of the variable**
- **μ = Mean**
- **σ = Standard deviation and σ^2 = Variance**
- **e ≈ 2.718 and π ≈ 3.142**

Normal distributions are important in statistics and are often used in the natural and social sciences to represent real-valued random variables whose distributions are not known. Their importance is partly due to the Central Limit Theorem. It states that, under some conditions, the average of many samples (observations) of a random variable with finite mean and variance is itself a random variable whose distribution converges to a normal distribution as the number of samples increases.

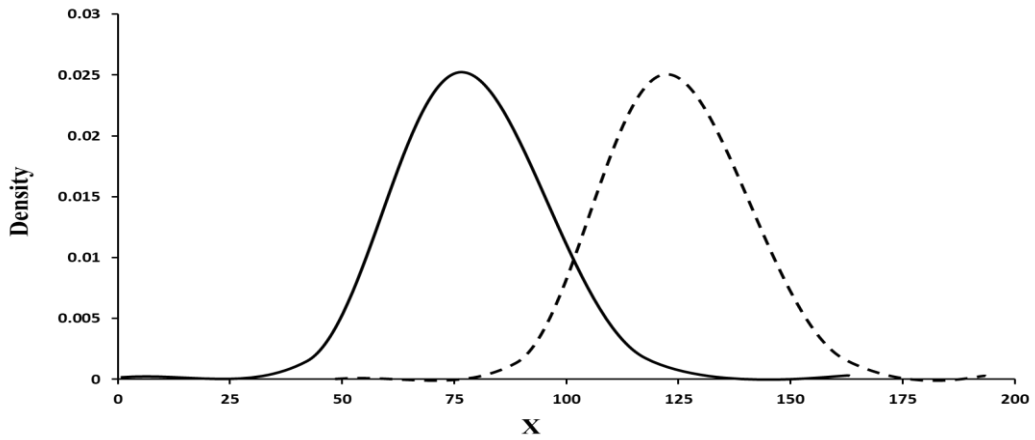
Mean and Variance of the Normal Distribution:

As with any probability distribution, the parameters for the normal distribution define its shape and probabilities entirely. The normal distribution has two parameters, the mean μ and standard deviation σ. The normal distribution does not have just one form. A random variable with a normal distribution is said to be normally distributed with mean **μ** and standard deviation **σ**. That is,

$$\mathbf{X \sim N(\mu\ ,\ \sigma)}$$

The Shape of the Normal Distribution:

The graph of the normal distribution depends on the two factors; the mean and the standard deviation. The mean of the distribution determines the location of the center of the graph, and the standard deviation determines the height and width of the graph. All normal distributions look like a symmetric, bell-shaped curve, as shown below. When the standard deviation is small, the curve is tall and narrow; and when the standard deviation is big, the curve is short and wide (see below).



Normal Distribution
Different Means – Same Standard Deviation

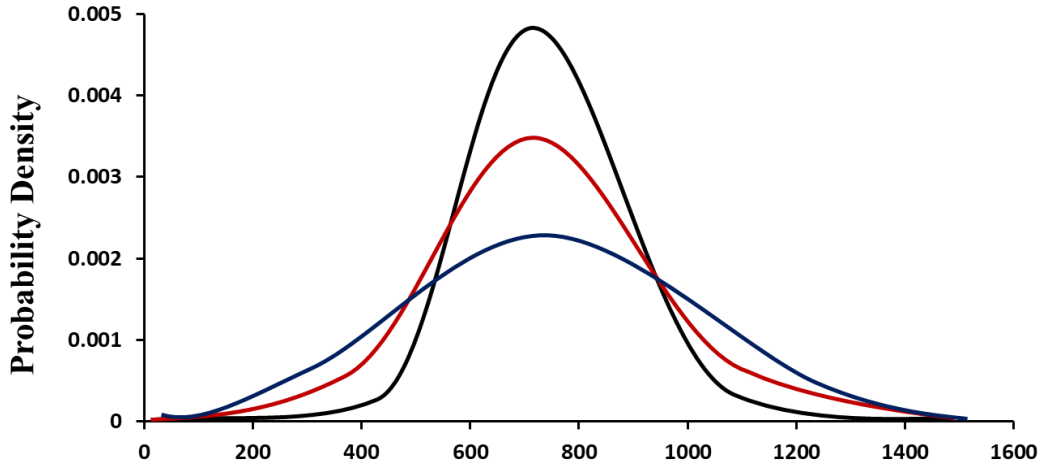
Mean:

The mean is the central tendency of the distribution. It defines the location of the peak for normal distributions. Most values cluster around the mean. On a graph, changing the mean shifts the entire curve left or right on the X-axis.

Standard Deviation:

The standard deviation is a measure of variability. It defines the width of the normal distribution. The standard deviation determines how far away from the mean the values tend to fall. It represents the typical distance between the observations and the average.

On a graph, changing the standard deviation either pull or spreads out the width of the distribution along the X-axis. Larger standard deviations produce distributions that are more spread out.



Normal Distribution With Different Standard Deviations

Why Do Normal Distributions Matter?

All kinds of variables in natural and social sciences are normally or approximately normally distributed. Height, birth weight, reading ability, job satisfaction, or scores are just a few examples of such variables.

Because normally distributed variables are so common, many statistical tests are designed for normally distributed populations.

Understanding the properties of normal distributions means you can use inferential statistics to compare different groups and make estimates about populations using samples.

Properties of Normal Distributions:

The scores or observations are most crowded (dense) in intervals around the mean, where the curve is highest. Towards the ends of the curve, the height is lower; the scores become less crowded the further from the mean we go. This tells us that observations around the mean are more likely to occur than observations

further from the center. In a random selection from the normal distribution, scores around the mean have a higher probability of being selected than scores far away from the mean.
The normal distribution is not really the normal distribution but a family of distributions.

Each of them has these properties:

- Unimodal (one mode)
- The total area under the curve is 1.
- The curve is symmetrical so that the mean, median and mode are located in the center.
- The curve is bell shaped (maximum height (mode) at the mean).
- The greatest proportion of scores lies close to the mean. The further from the mean one goes (in either direction) the fewer the scores.
- Asymptotic (the further the curve goes from the mean, the closer it gets to the X axis; but the curve never touches the X axis).

Area Under Normal Curve:

Empirical Rule:

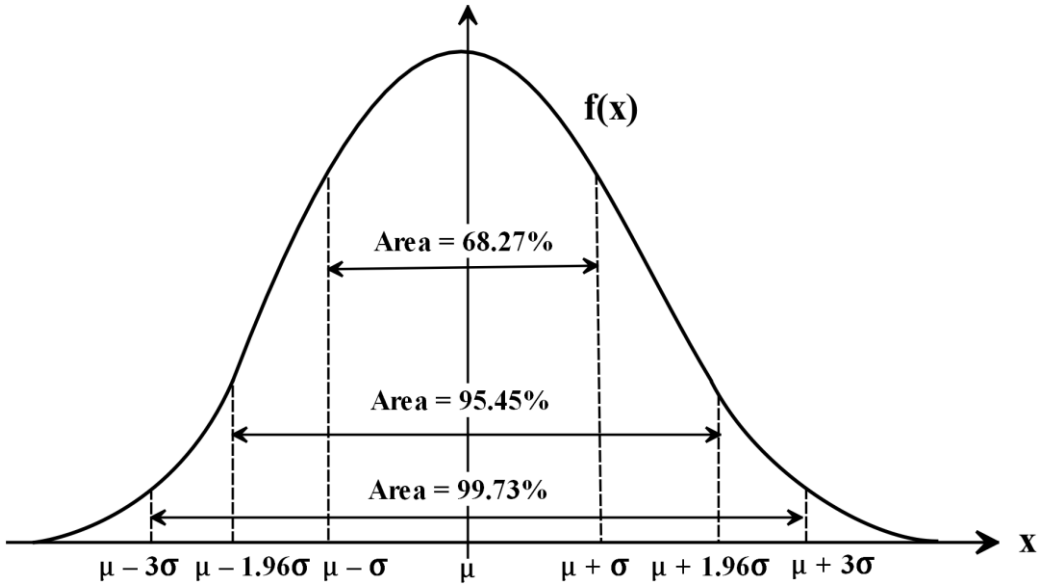
The **empirical rule**, or the **68% - 95% - 99.7% Rule**, tells you where most of your values lie in a normal distribution:

- Around **68%** of values are within **1 standard deviation** from the mean.
- Around **95%** of values are within **1.96 standard deviations** from the mean.
- Around **99.7%** of values are within **3 standard deviations** from the mean.

For example: Suppose that the STAT scores of a group of students follow a normal distribution with a mean score (**μ**) of 1150 and a standard deviation (**σ**) of 150.

Following the empirical rule**:**

- Around 68% of scores are between 1000 and 1300, one standard deviation above and below the mean.
- Around 95% of scores are between 850 and 1450, 1.96 standard deviations above and below the mean.
- Around 99.7% of scores are between 700 and 1600, 3 standard deviations above and below the mean.



The empirical rule is a quick way to get an overview of your data and check for any outliers or extreme values that don't follow this pattern.

If data from small samples do not closely follow this pattern, then other distributions like the **t - distribution** may be more appropriate. This distribution is to be considered later. Once you identify the distribution of your variable, you can apply appropriate statistical tests.

3.3 The Standard Normal Distribution

The **standard normal distribution**, also called the **z-distribution**, is a special normal distribution where the mean is 0 and the standard deviation is 1. Every normal distribution is a version of the standard normal distribution that's been stretched or squeezed and moved horizontally right or left. While individual observations from normal distributions are referred to as x, they are referred to as z in the z-distribution. Every normal distribution can be converted to the standard normal distribution by turning the individual values into z-scores. Z-scores tell you how many standard deviations away from the mean each value lies.

You only need to know the mean and standard deviation of your distribution to find the *z*-score of a value.

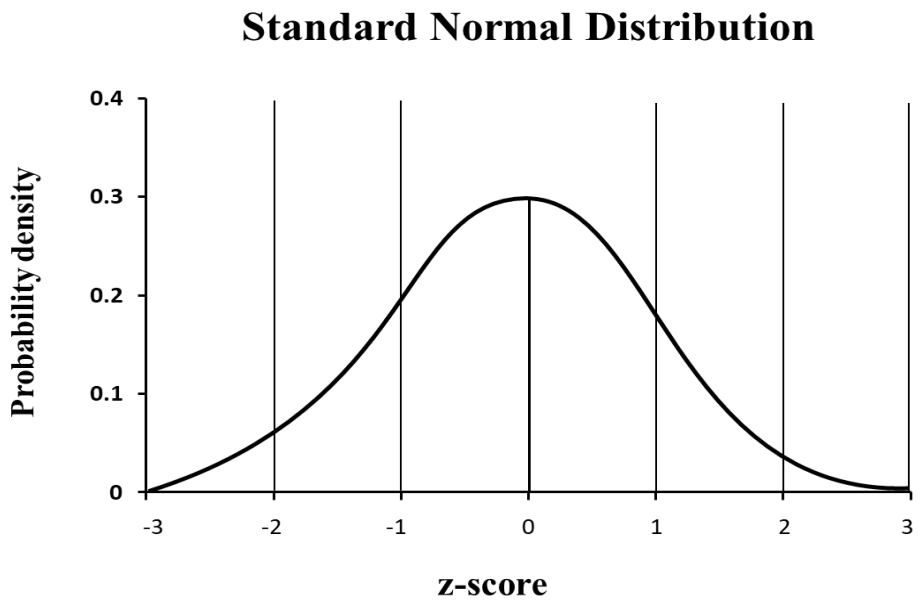
Z-Score Formula	**Explanation**
$\mathbf{Z = \frac{x - \mu}{\sigma}}$	**x = individual value** **μ = mean** **σ = standard deviation**

We convert normal distributions into the standard normal distribution for several reasons**:**

- To find the probability of observations in a distribution falling above or below a given value.
- To find the probability that a sample mean significantly differs from a known population mean.
- To compare scores on different distributions with different means and standard deviations.

Finding Probability Using z- distribution:

Each z-score is associated with a probability, that tells you the probability of values below that z-score occurring. If you convert an individual value into a z-score, you can then find the probability



of all values up to that value occurring in a normal distribution using the area under the standard normal curve as shown in the following example.

Example (3.7):

Suppose that the annual salaries of employees in a large company is normally distributed with mean $\mu = \$60{,}000$ and standard deviation $\sigma = \$15{,}000$.

(a) Find the probability of a randomly selected employee earning less than \$75,000 annually.

(b) Find the probability of randomly selected employee that earns more than \$80,000. What percent of employees that earn more than this salary?

(c) Find the range of annual salaries of the top 15% earners.

Solution:

Hint: The solution of this example will only be presented with a detailed explanation of the steps needed to solve it. For the rest

of examples, we will follow the same steps without detailed explanation of each step.

For purposes of simplicity, let us express the salaries in thousands of dollars.

(a) To answer this question, we have to find the portion of the area under the normal curve from 75 all the way to the left.

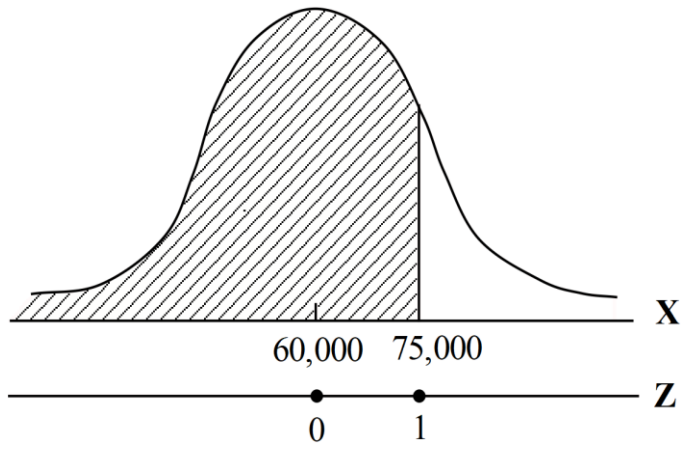
To find this area, you can apply complex mathematical formulas, or you can use the Z – table (represents the area under Standard Normal curve), in which statisticians have already applied those formulas for you. Because they could not develop tables for every possible combination of the mean and the standard deviation, statisticians developed one standardized and simplified normal distribution with the mean of 0 and the standard deviation of 1.

All other distributions with different μ and σ can be converted into a standardized normal distribution using the transformation formula:

$$\mathbf{Z = \frac{x - \mu}{\sigma}}$$

Find the value of Z and then find the number that corresponds to that Z in the body of Z-table**:**

$$Z = \frac{75 - 60}{15} = 1$$



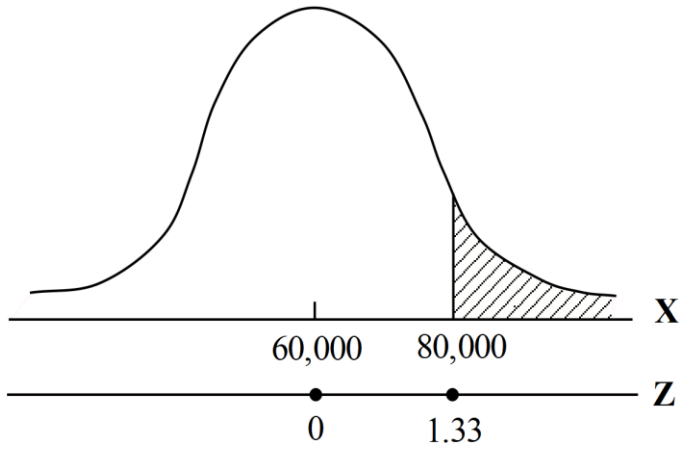
The table value that corresponds to –1 is 0.1587. That is,

$$P(X \leq 75) = 0.8413$$

This number indicates the area under the curve from 45 all the way to the left. It also indicates that the probability of randomly selecting an employee who makes less than $45000 a year is 0.8413.

(b) Find the Z - value using the transformation formula**:**

$$Z = \frac{80 - 60}{15} = 1.33$$

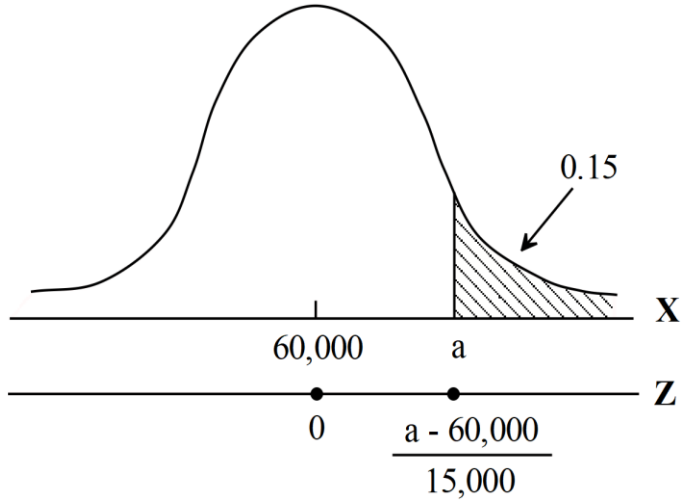


The table value that corresponds to Z = 1.33 is 0.9082. This is not the final answer, however, because as you can see, the Z-table only shows the values less than (and to the left of) each value of Z. 0.9082 that we found in the table is the probability of randomly selecting an employee that earns less than $80000 annually. If you remember that the entire normal curve covers 100% of the distribution, you will be able to find the complement probability or the area under the curve to the right of 80. Just subtract the table value from 1, That is,

$$P(X > 80) = 1 - P(X \leq 80)$$
$$= 1 - 0.9082 = 0.0918$$

If we multiply the value of the probability by 100, we obtain percentage. Thus, the required percentage is 100(0.018) = 1.8%.

(c)



In this situation, the situation is reversed. You are given an area under the Standard Normal curve and have to find a specific value of x that would correspond to the annual salary that separates the highest-paid 15% of employees. In this situation, you have to move backwards. First, find 80% (or 0.85) in the body of the Z-table. Again, you use 85% (100% - 15%) because Z-table only works for the area below each value of Z. Go across and up to the margins of the table to find the value of Z. That value is 1.04. Now, you have all the data to use the transformation formula and solve for x**:**

$$Z = \frac{X - \mu}{\sigma} \quad \text{giving} \quad 1.04 = \frac{X - 60}{15}$$

$1.04 \times 15 = x - 60$

$X = 15.6 + 60 = 75.6$

Thus, the highest-paid 15% of employees make $75,600 and above per year.

Example (3.8):

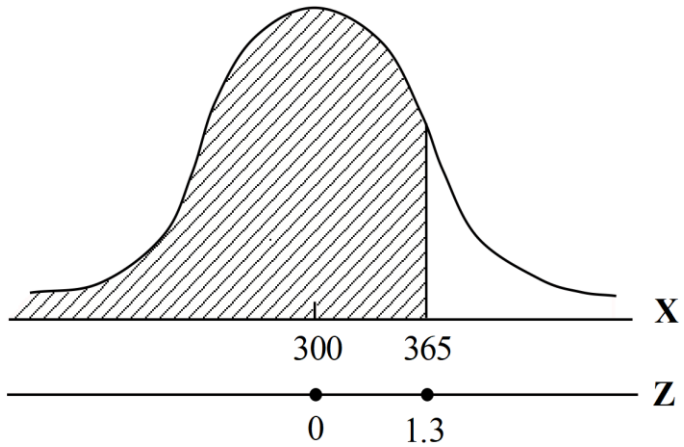
An average light bulb manufactured by a corporation lasts 300 days with a standard deviation of 50 days. Assuming that bulb life is normally distributed, what is the probability that a light bulb will last at most 365 days?

Solution:

Given a mean score of 300 days and a standard deviation of 50 days, we want to find the cumulative probability that bulb life is less than or equal to 365 days. Thus, we know the following**:**

- The value of the normal random variable is 365 days.
- The mean is equal to 300 days.
- The standard deviation is equal to 50 days.

$$P(X \leq 365) = P\left(\frac{X - 300}{50} \leq \frac{365 - 300}{50}\right)$$
$$= P(Z \leq 1.3)$$



The table value that corresponds to Z = 1.3 is 0.9032
Hence, there is a 90.32% chance that a light bulb will burn out within 365 days.

Example (3.9):

Suppose scores on an IQ test are normally distributed. If the test has a mean of 100 and a standard deviation of 10, what is the probability that a person who takes the test will score between 90 and 110?

Solution:

Here, we want to know the probability that the test score falls between 90 and 110. The "trick" to solving this problem is to realize the following**:**

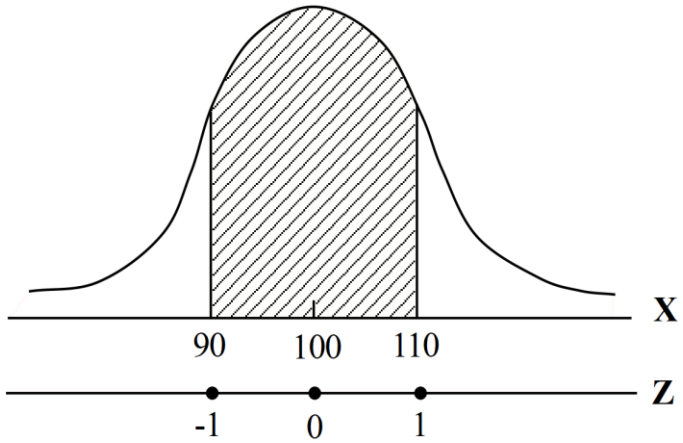
$$P(90 < X < 110) = P(X < 110) - P(X < 90)$$

$$P(X \leq 110) = P\left(\frac{X - 100}{10} \leq \frac{110 - 100}{10}\right) = P(Z \leq 1)$$

The table value that corresponds to Z = 1 is 0.8413

Similarly, $P(X \leq 90) = P\left(\frac{X - 100}{10} \leq \frac{90 - 300}{10}\right)$

$= P(Z \leq -1)$



Note that: $P(Z \leq -1) = 1 - P(Z \leq 1)$ because Z- curve is symmetric.

$= 1 - 0.8413 = 0.1587$

We use these findings to compute our final answer as follows:

$$P(90 < X < 110) = P(X < 110) - P(X < 90)$$

$$= 0.8413 - 0.1587 = 0.6826$$

Thus, about 68.26% of the test scores will fall between 90 and 110.

Example (3.10):

Most graduate schools of business require applicants for admission to take the Graduate Management Admission Council's GMAT examination. Scores on the GMAT are roughly normally distributed with a mean of 527 and a standard deviation of 112.

(a) What is the probability of an individual scoring above 500 on the GMAT?

(b) How high must an individual score on the GMAT in order to score in the highest 5%?

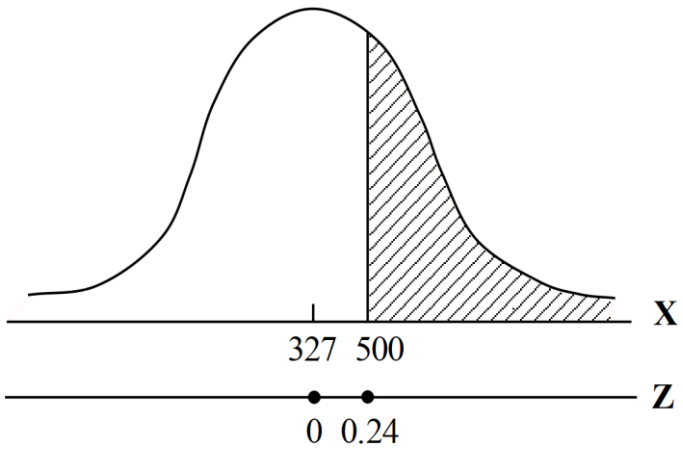
Solution:

(a) $P(X > 500) = 1 - P(x \leq 500)$

$$P(x \leq 500) = P\left(\frac{X - 527}{112} \leq \frac{500 - 527}{112}\right)$$

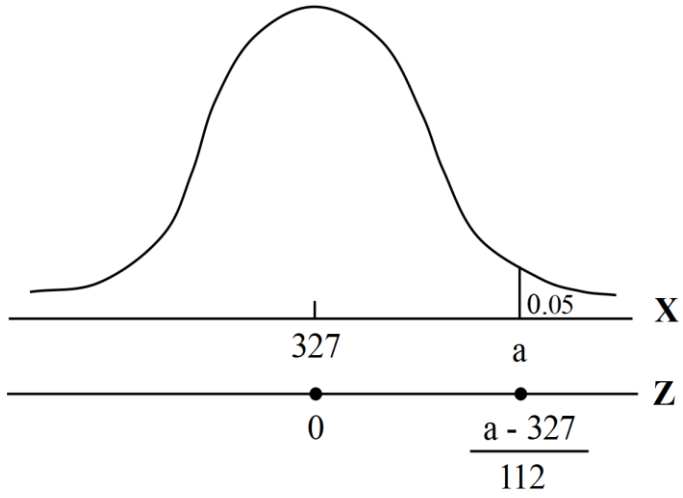
$$= P(Z \leq -0.24) = 1 - P(Z \leq 0.24)$$

$$= 1 - 0.5948 = 0.4052$$



Thus, $P(X > 500) = 1 - 0.4052 = 0.5948$

(b)



μ = 527 and σ = 112

P(X > a) = 0.05

P(X > a) = 1 - P(X ≤ a) = 1 – 0.5 = 0.95

$$P(x \le a) = P\left(\frac{a - 527}{112} \le \frac{500 - 527}{112}\right)$$

$$= P\left(\frac{a - 527}{112} \le -0.24\right) = 0.95$$

This gives Z = 1.645 From the Z-table

$$\frac{a - 527}{112} = 1.645$$

a = 527 + 1.645(112) = 711.24

Thus, to score in the highest 5% you should have at least the score 711.24.

Example (3.11):

Entry to a certain University is determined by a national test. The scores on this test are normally distributed with a mean of 500 and a standard deviation of 100. A student wants to be admitted to this university and he knows that he must score better than at least 70% of the students who took the test. He takes the test and scores 585. Will he be admitted to this university?

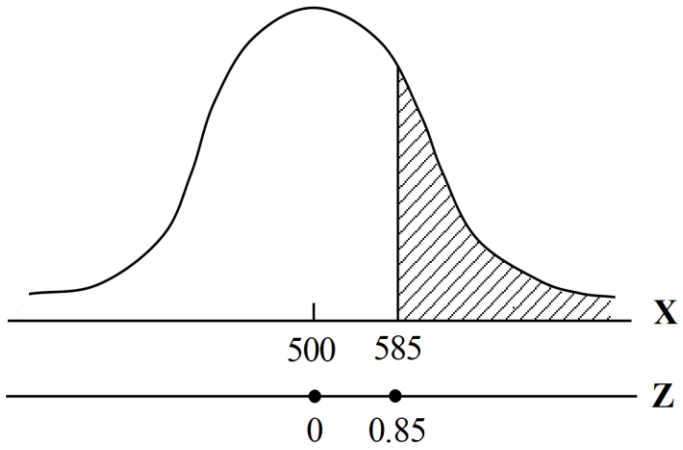
Solution:

Let x be the random variable that represents the scores. x is normally distributed with a mean of 500 and a standard deviation of 100. The total area under the normal curve represents the total number of students who took the test.

$\mu = 500$, $\sigma = 100$ and $x = 585$

$$P(X \leq 385) = P\left(\frac{X - 500}{100} \leq \frac{585 - 500}{100}\right)$$

$$= P(Z \leq 0.85)$$



The proportion P of students who scored below 585 is given by**:**

P[area to the left of z = 0.85] = 0.8023

If we multiply the values of the areas under the curve by 100, we obtain percentages.

Thus, the percentage of students who scored below 585 is 80.23% This student scored better than 80.23% of the students who took the test and he will be admitted to this University.

Example (3.12):

A radar unit is used to measure speeds of cars on a motorway. The speeds are normally distributed with a mean of 90 km/hr and a standard deviation of 10 km/hr. What is the probability that a car picked at random is travelling at more than 100 km/hr?

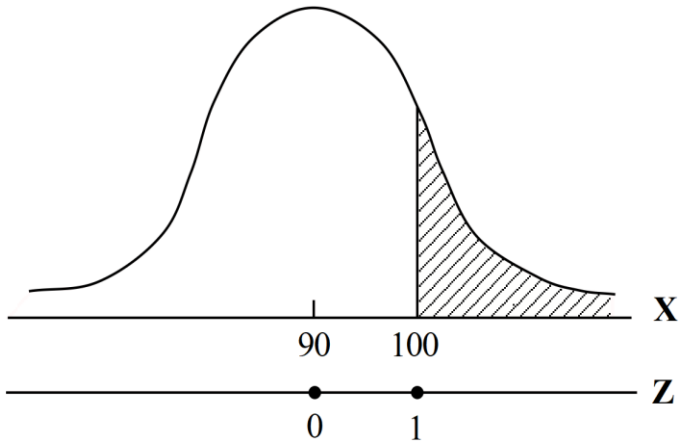
Solution:

Let x be the random variable that represents the speed of cars. The variable x has μ = 90 and σ = 10. We have to find the probability that x is higher than 100 or P(x > 100).

$$P(x > 100) = 1 - P(x \geq 100)$$

$$P(x \leq 100) = P\left(\frac{X - 90}{10} \leq \frac{100 - 90}{10}\right)$$

$$= P(Z \leq 1) = 0.8413$$



P(x > 100) = P(z > 1) = [total area] - [area to the left of z = 1]
= 1 - 0.8413 = 0.1587

The probability that a car selected at a random has a speed greater than 100 km/hr is equal to 0.1587.

Exercises for Sections 3.2 & 3.3
(Exam Questions)

(3.8) The lifetime of a certain kind of battery is normally distributed with a mean of 300 hours and a standard deviation of 50 hours.

(a) Find the probability that the lifetime of a battery randomly selected is**:**

(i) between 202 and 429 hours.

(ii) more than 429 hours.

(b) Determine the value of the lifetime above which the best 5% of the batteries will lie.

(c) What are the **MINITAB** commands that are required for answering the two **Parts (a)** and **(b)**.

(Exam 1998)

(3.9) If the Intelligence Quotient (I.Q.) scores of a population of individuals is approximately normally distributed with a mean of 100 and a standard deviation of 16.

(1) What is the probability that a randomly selected individual will have an I.Q.

(a) Less than or equals 80.

(b) Between 90 and 110.

(2) What I.Q. score is associated with the lowest 5% of individuals?

(3) What interval of scores, centered at 100, include 50% of individuals?

(4) For a sample of size 64 drawn from this population, what is the probability of obtaining a sample mean of 104 at most?

(5) Use **MINITAB** to answer **Parts (a) to (d).**

(Exam 1999)

(3.10) The scores of statistics for a group of students were approximately normally distributed with a mean of 4 and a standard deviation of 2. If a student is randomly selected from among this group.

(a) find the probabilities that he will have the following scores**:**

(i) Above 15 **(ii)** Between 12 and 16

(b) Use **MINITAB** to answer Part (a).

(3.11) If the weights of students of one college are normally distributed with mean 65 kg, and standard deviation σ . If 67% of the students' weights are less than 70 kg, find the value of σ.

(Exam 2001)

(3.12) The time taken by new employees at a corporation to learn a packing procedure is normally distributed with a mean of **24** hours and a standard deviation of **2.5** hours.

(a) What percentage of new employees can learn this procedure in more than **20** hours**?**

(b) Find the probability that a randomly selected new employee will take **26** to **30** hours to learn this procedure.

(c) Find the time over which only **20%** of the new employees take to learn this procedure.

(3.13) Hourly wage rates for unskilled workers in a particular nationwide industry are normally distributed with a mean of **$5** and a standard deviation of **$0.8**. An employee is selected at random.

(a) Find the probability that he will earn a basic rate of**:**

(i) between **$3** and **$7** per hour **(ii)** more than **$6** per hour.

(b) Approximately **80%** earn more than the recommended minimum basic rate. What is this minimum rate**?**

(c) Use **MINITAB** to answer **Parts (a)** and **(b)**. **(Exam 2002)**

(3.14) The final examination scores in statistics are normally distributed with a mean of **70** and a standard deviation of **10**.

(a) Find the probability that the score of a randomly selected student is**:**

(i) at most **70** **(ii)** between **50** and **90**

(b) What should the score be so that only **5%** of the students get a score higher than this**?**

(c) If you achieved a score of **85**, how far, in standard deviations, did your score depart from the mean**?** What percentage of those who took the examination scored higher than you**?**

(Exam 2003 Qena)

(3.15) A certain brand of light bulbs has a lifetime which is normally distributed with a mean of **100** hours and a standard deviation of **8** hours.

(a) Find the probability that the lifetime of a randomly selected bulb is**:**

(i) At most **90** hours **(ii)** More than **120** hours

(b) What percentage of bulbs could be expected to have a lifetime of more than **100** hours**?** To what value must the standard deviation be reduced in order to **halve** this percentage**?**

(c) Determine the value of the lifetime above which the best **5%** of the bulbs will lie.

(Exam 2003 Sohag)

(3.16) A firm's marketing manager believes that the total sales for the firm next year can be modeled by using a normal distribution with a mean of **$2.5 million** and a standard deviation of **$300,000**.

(1) What is the probability that the firm's sales will exceed **$3 million?**

(2) What is the probability that the firm's sales will fall within **$150,000** of the **expected** level of sales**?**

(3) In order to cover fixed costs, the firm's sales must **exceed** the break-even level of **$1.8 million**. What is the probability that the firm will cover fixed costs**?**

(4) Determine the sales level that has only a **5%** chance of being **exceeded** next year.

(Exam 2004)

(3.17) The average weekly pay for a production worker in a certain occupation is L.E **1000**. Assume that wages are **normally** distributed with a standard deviation of L.E. **80**.

(1) For a randomly selected production worker, find the probability of his weekly wage being**:**

(a) L.E. **900** or less.

(b) Within L.E **800** to L.E **1200**.

(2) How much does a production worker have to make to be in the top **60%** of wage earners**?**

(Exam 2005)

(3.18) A study of Furniture Wholesales, Inc. regarding the payment of invoices revealed that, on the average, an invoice was paid **20** days after it was received. The standard deviation

(Exam 2006)↓

equaled **5** days. Assuming the payments are normally distributed**:**

(1) What **percent** of the invoices are paid within **15** days of receipt**?**

(2) What is the probability of selecting any invoice and finding it was paid between **18** and **26** days after it was received**?**

(3) The management of Furniture Wholesales wants to encourage their customers to pay their monthly invoices as soon as possible. Therefore, it was announced that a 2 percent reduction in price would be in effect for customers who pay within **7** working days of the receipt of the invoice. Out of **200** customers during July, how many would normally be eligible for the reduction**?**

(4) What are the limits that contain **95%** of the payments of invoices equally around the mean**?**

(Exam 2006)

(3.19) The time (x) needed to complete a final examination in a particular collage is normally distributed with a mean of 80 minutes and a standard deviation of 10 minutes.

(1) What is the probability of completing the exam in one hour or less?

(2) What is the probability that a student will complete the exam in more than 60 minutes but less than 75 minutes?

(3) Assume that the class has 60 students and that the examination period is 90 minutes in length. How many students do you expect will be unable to complete the exam in the allotted time?

(4) Two students are chosen at random from this collage. What is the probability that at least one of them takes one hour or less to complete this exam?

(Exam 2007)↓

(5) Suppose the mean and standard deviation are unknown, and $P(x \leq 70) = P(x > 94) = 0.0668$. Find the mean and standard deviation.

(Exam 2007)

(3.20) The amount of time devoted to studying statistics each week by students who achieve a grade of excellent in the course is a **normally** distributed random variable with a mean of **7.5** hours and a standard deviation of **1.8** hours.

(1) What **proportion** of excellent students spend more than **9** hours studying**?**

(2) Find the probability that an excellent student spends between **6** and **9** hours studying.

(3) What is the amount of time below which only **10%** of all excellent students spend studying**?**

(3.21) The annual sales of a product are **normally** distributed with an unknown mean and an unknown standard deviation. **40%** of the sales are more than **470,000**, and **10%** of the sales are more than **500,000**. What are the **mean** and the **standard deviation?**

(Exam 2008)

(3.22) A manufacturer finds that although he promises delivery of a certain item in **7** weeks, the time he takes to deliver to customers is approximately normally distributed with a mean of **6** weeks and a standard deviation of **2** weeks.

(1) What **proportion (probability)** of customers receive their deliveries late**?**

(2) What **Percentage** of customers receive their deliveries within **4** to **7** weeks**?**

(Exam 2009)↓

(3) To what figure should his delivery promise be altered if it is required that only **20%** of deliveries should be late**?**

(4) Find the probability that customers will receive deliveries within **5** weeks if the manufacturer reduces the standard deviation of delivery time to **one** week, keeping the mean time at **6** weeks**?**

(Exam 2009)

(3.23) The final examination scores in statistics are **normally** distributed with a **mean** of **70** and a **standard deviation** of **10**.

(1) Find the probability that the score of a randomly selected student is

(a) at most 85. **(b) greater than 60**.

(2) What score is associated with the **highest 25%** of students**?**

(3) Determine the **interquartile range** ($Q_3 - Q_1$) of the distribution.

(Exam 2013)

(3.24) In a factory, the daily wages **(X)** for production workers follow a **normal** distribution with a **mean** of L.E.**100** and a **standard deviation** of L.E.**10**.

(1) What is the probability that the wage of a randomly selected worker will be**:**

(a) L.E. 120 or less. **(b) greater than L.E.72**.

(2) How much does a production worker have to make to be in the **top 10%** of wage earners**?**

(3) Between what **two** wages **(A** and **B) symmetrically** distributed around the mean will **95%** of the workers fall**?**
i.e. $\mathbf{P(X > B) = P(X \leq A)}$.

(Exam 2014)

(3.25) A set of final examination grades in an applied statistics course was found to be **normally** distributed with a **mean** of **73** and a **standard deviation** of **8**.

(1) What is the **probability** of getting a grade of **81 or less** on this exam?

(2) Only **20%** of the students taking the test scored **higher than** what grade**?**

(3) If the professor gives A's (Excellent) to the top **10%** of the class regardless of the score, are you **better off** with a grade of **81** on this exam or a grade of **68** on a different exam where the **mean** is **62** and the **standard deviation** is **3**? Show your answer statistically and **explain**.

(Exam 2015)

(3.26) Anticipated consumer demand for a product next month can be represented by a **normal** random variable with **mean 1,200** units and **standard deviation 100** units.

(1) What is the **probability** that **demand** will **exceed 1,300** units?

(2) Find the probability that **demand** will be **between 1,000** and **1,300** units?

(3) The **probability** is **0.9** that **demand** will be **more than** how many units?

(Exam 2016)

(3.27) The marks **(x)** of the large number of candidates taking a particular examination were such that it was reasonable to assume that these marks were normally distributed. **Half** the candidates in the examination scored **more than 45** marks and **25%** of the candidates scored **more than 51.7** marks.

(Exam 2017)↓

That is, **P(x > 45) = 0.5** and **P(x > 51.7) = 0.25**.

(1) Determine the **mean** and **variance** of the distribution.

(2) The **bottom 30%** of the candidates **failed**. Determine the **pass mark** for the examination.

(3) **Based on P(x > 51.7 = 0.25)**, determine the **first quartile (Q_1)** of the distribution.

(Exam 2017)

(3.28) Anticipated consumer demand **(x)** for a product next month can be represented by a **normal** random variable with **mean 1200** units and **standard deviation 100** units.

(1) What is the probability that demand is **1300** units **or less**?

(2) The probability is **0.025** that demand will be **more than how many** units?

(3) Between what two amounts of demand (**A** and **B**) **symmetrically** distributed around the mean will **95%** of sales fall? **i.e. P(x ≤ A) = P(x > B)**.

(Exam 2018)

(3.29) **Area** under the **normal** curve on **either side** of mean is

(A) 0.5 (B) 1 (C) Mean value (D) 2

(3.31) The **median** of the **standard normal** distribution equal to

(A) μ (B) 0 (C) 1 (D) 0.5

(3.32) The working life of a certain type of light bulb is **normally** distributed with a **mean** of **500** hours and a **standard deviation** of **60** hours.

(a) What is the **probability** that a bulb works **at most 560** hours?

(A) 0.8413 (B) 0.1578 (C) 0.8422 (D) 0.1587

(MCQ Exam 2019)↓

(b) If a bulb is still working **after 440** hours of operation, what is the **probability** that its lifetime **exceeds 560** hours?

(A) 0.1688 (B) 0.1886 (C) 0.1868 (D) 0.1685

(MCQ Exam 2019)

(3.30) Weekly wage rates for unskilled workers in a particular nationwide industry are normally distributed with a mean of L.E. 60 and a standard deviation of L.E. 5.

(a) Find the probability that an employee selected at random will earn a basic rate of between L.E. 65 and L.E. 75.

(b) In a group of 400 unskilled employee, how many would you expect to earn more than L.E. 60?

(c) Approximately 20% earn less than the recommended minimum basic rate. What is the minimum rate?

Note: This Question is not in the right order.

(Exam 2000)

3.4 t - Distribution:

Why Use the t - Distribution?

According to the central limit theorem, the sampling distribution of a statistic (like a sample mean) will follow a normal distribution, as long as the sample size is sufficiently large. Therefore, when we know the standard deviation of the population, we can compute a z-score, and use the normal distribution to evaluate probabilities with the sample mean.

But sample sizes are sometimes small, and often we do not know the standard deviation of the population. When either of these problems occur, statisticians rely on the distribution of the **t statistic** (also known as the **t - score**), whose values are given by**:**

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

where,

$\bar{x}$ is the sample mean,

μ is the population mean,

s is the standard deviation of the sample, and

n is the sample size.

The distribution of the t - statistic is called the **t - distribution** or the **Student t - distribution**.

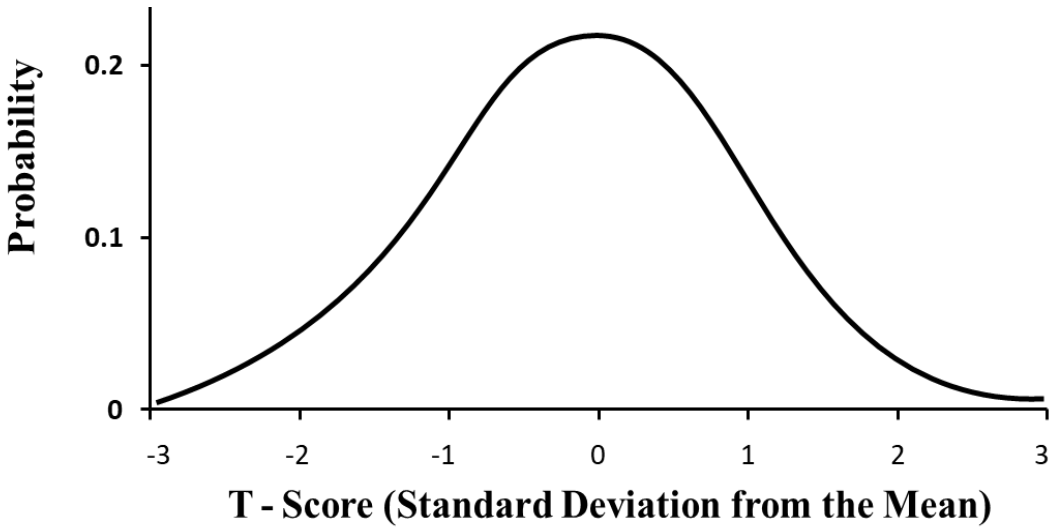
The t- distribution allows us to conduct statistical analyses on certain data sets that are not appropriate for analysis, using the normal distribution.

Properties of the t - distribution:

The t distribution has the following properties**:**

- The mean of the distribution is equal to 0.

- The shape of the t – distribution is symmetric around the mean.



- The variance is equal to **v/(v - 2)**, where **v** is the degrees of freedom and **v ≥ 2,** where v = n - 1.
- The variance is always greater than 1, although it is close to 1 when there are many degrees of freedom. With infinite degrees of freedom, the t-distribution is the same as the standard normal distribution.

Degrees of Freedom:

There is actually many different t-distributions. The particular form of the t-distribution is determined by its **degrees of freedom**. The degree of freedom refers to the number of independent observations in a set of data.

When estimating a mean score or a proportion from a single sample, the number of independent observations is equal to the sample size minus one. Hence, the distribution of the t - statistic from samples of size **8** would be described by a t-distribution having 8 - 1 or **7 degrees of freedom**. Similarly, a t-distribution

having 15 degrees of freedom would be used with a sample of size 16. The degrees of freedom will change depending on the statistic being used.

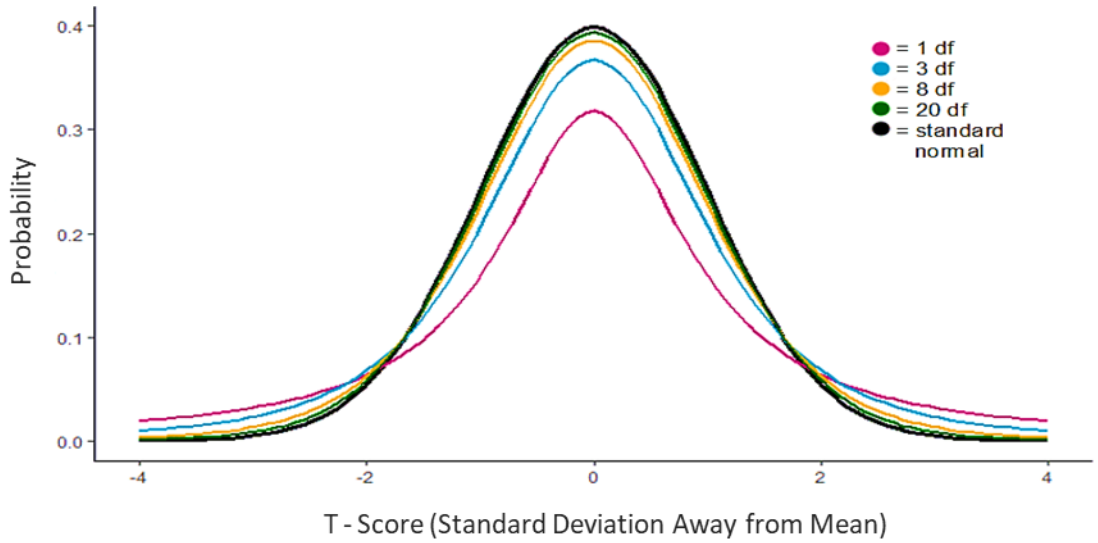
For other applications, the degrees of freedom may be calculated differently. We will describe those computations as they come up.

As the degrees of freedom (total number of observations minus 1) increases, the t-distribution will get closer and closer to matching the standard normal distribution, a.k.a. the z-distribution, until they are almost identical.

Above 30 degrees of freedom, the t-distribution roughly matches the z-distribution. Therefore, the z-distribution can be used in place of the t-distribution with large sample sizes.

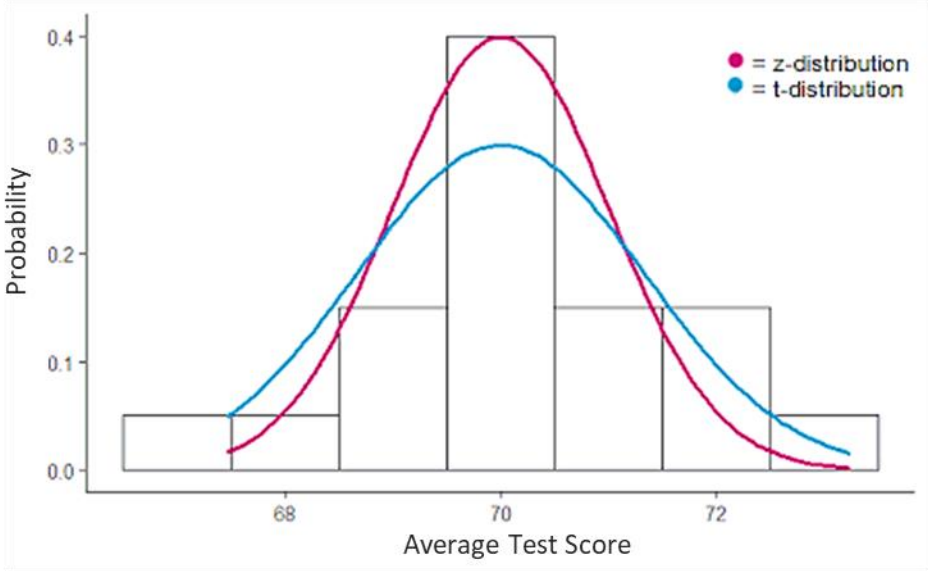
The z-distribution is preferable over the t-distribution when it comes to making statistical estimates because it has a known variance. It can make more precise estimates than the t-distribution, whose variance is approximated using the degrees of freedom of the data.

The t-distribution should not be used with small samples from populations that are not approximately normal.



T- distribution and the Standard Normal Distribution:

The **Student's t** distribution is very similar to the standard normal distribution.



- It is symmetric about its mean.
- It has a mean of zero.
- It has a standard deviation and variance greater than 1.
- There are actually many t distributions, one for each degree of freedom.
- As the sample size increases, the t distribution approaches the normal distribution.
- It is bell shaped.
- The t-scores can be negative or positive, but the probabilities are always positive.

Probability and the Student t - Distribution:

When a sample of size n is drawn from a population having a normal (or nearly normal) distribution, the sample mean can be transformed into a t - statistic, using the equation presented at the beginning of this section. We repeat that equation below**:**

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

Where $\bar{x}$ is the sample mean, **μ i**s the population mean, **S** is the standard deviation of the sample, **n** is the sample size, and degrees of freedom are equal to **n - 1**.

The t- statistic produced by this transformation can be associated with a unique **cumulative probability**. This cumulative probability represents the probability of finding a sample mean less than or equal to x, given a random sample of size n.

Example (3.13):

The chief executive officer (CEO) of light bulbs manufacturing company claims that an average light bulb lasts 300 days. A researcher randomly selects 24 bulbs for testing. The sampled bulbs last an average of 275 days, with a standard deviation of 49 days. If the CEO's claim were true, what is the probability that 24 randomly selected bulbs would have an average life of no more than 275 days?

Solution:

The traditional approach requires you to compute the t - statistic, based on data presented in the problem description.

The first thing we need to do is compute the t - statistic, based on the following equation**:**

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

Using the formula**:**

$$t = \frac{275 - 300}{49/\sqrt{24}} = -2.5$$

The degrees of freedom are equal to 24 – 1 = 23

Assuming the CEO's claim is true, the population mean equals 300.
The sample mean equals 275.
The standard deviation of the sample is 49.

$$P(\bar{x} \leq 275) = P\left(\frac{\bar{x} - 300}{\frac{49}{\sqrt{24}}} \leq \frac{275 - 300}{\frac{49}{\sqrt{24}}}\right)$$

$$= P(t \leq -2.5) = P(t > 2.5) \quad \text{symmetric distribution}$$

From the t-table given in the Appendix at the end of this book, the cumulative probability is 0.01, that is,

$$P(x \leq 275) = 0.01$$

Hence, if the true bulb life were 300 days, there is a 1% chance that the average bulb life for 24 randomly selected bulbs would be less than or equal to 275 days.

Example (3.14):

For a t- distribution with **8** degrees of freedom, find the following probabilities:

(a) $P(t \leq 2.869)$ (b) $P(t \leq 0.306)$ (c) $P(t > 1.397)$

Solution:

From the t-table given in the Appendix at the end of this book, we find:

(a) $P(t \leq 2.869) = 1 - P(t > 2.869) = 1 - 0.01 = 0.99$

(b) $P(t \leq 0.703) = 1 - P(t > 0.306) = 1 - 0.025 = 0.975$

(c) $P(t > 1.397) = 0.1$

Exercises for Section 3.4

(3.33) Suppose you are designing a test. The scores on an that test are normally distributed, with a population mean of 60. Suppose 25 people are randomly selected and tested. The standard deviation in the sample group is 10. What is the probability that the average test score in the sample group will be at most 64.13?

(3.34) Suppose you are hired to test the quality of products. The company claims that an average usage lasts 100 days. You randomly select 20 products for testing. The sampled products last an average of 90 days, with a standard deviation of 16 days. Then you can try to calculate the probability of having no more than an average of 87 days to see if the company claim were true.

(3.35) Find the value for t for each of the following questions.

(a) What is the value of t which has an area to its right of 0.05 when the sample size is n = 28.

(b) What is the value of t which has an area to its right of 0.025 when the sample size is n = 41.

(c) What is the value of t which has an area to its right of 0.500 when the sample size is n = 11.

(d) What is the value of t which has an area to its left of 0.01 when the sample size is n = 76.

(e) What is the value of t which has an area to its left of 0.10 when the sample size is n = 33.

(3.36) Find the following probabilities using the t distribution with n = 23.

(a) What is the probability that *t* will be greater than or equal to 0? That is, what is $p(t \geq 0, df = 23 - 1 = 22)$?

(b) What is the probability that t will be greater than or equal to 2.074? That is, what is p(t ≥ 2.074, d*f* = 23 – 1 = 22)?

(c) What is the probability that t will be less than or equal to 2.819? That is, what is p(t ≤ 2.819, df = 23 – 1 = 22)?

(d) What is the probability that t will be less than or equal to -1.321? That is, what is p(t ≤ 1.321, df = 23 – 1 = 22)?

(e) What is the probability t will fall between -2.819 and 1.717? That is, what is the p(–2.819 ≤ t ≤ 1.717)?

Chapter (4)
Sampling Distributions

Contents

Chapter 4

Sampling Distributions

4.1 Sampling Distribution and Inferential Statistics:

Sampling distributions are important for inferential statistics. In practice, one will collect sample data and, from these data, estimate parameters of the population distribution. Thus, knowledge of the sampling distribution can be very useful in making inferences about the overall population.

Inferential statistics involves generalizing from a sample to a population. A critical part of inferential statistics involves determining how far sample statistics are likely to vary from each other and from the population parameter. These determinations are based on sampling distributions. The sampling distribution of a statistic is the distribution of that statistic, considered as a random variable, when derived from a random sample of size n. It may be considered as the distribution of the statistic for all possible samples from the same population of a given size. Sampling distributions allow analytical considerations to be based on the sampling distribution of a statistic rather than on the joint probability distribution of all the individual sample values.

The sampling distribution depends on the following**:**

- The underlying distribution of the population,
- The statistic being considered,
- The sampling procedure employed,
- The sample size used.

More Properties of Sampling Distributions:

1. The overall shape of the distribution is symmetric and approximately normal.

2. The center of the distribution is very close to the true population mean.

Finally, the variability of a statistic is described by the spread of its sampling distribution. This spread is determined by the sampling design and the size of the sample. Larger samples give smaller spread. As long as the population is much larger than the sample (at least 10 times as large), the spread of the sampling distribution is approximately the same for any population size

4.2 Sampling Distribution of Sample Mean:

For example, knowing the degree to which means from different samples differ from each other and from the population mean would give you a sense of how close your particular sample mean is likely to be to the population mean. Fortunately, this information is directly available from a sampling distribution. The most common measure of how much sample means differ from each other is the standard deviation of the sampling distribution of the mean. This standard deviation is called the **standard error of the mean**.

Suppose you randomly sampled 40 women between the ages of 21 and 35 years from the population of women in city, and then computed the mean height of your sample. You would not expect your sample mean to be equal to the mean of all women in this city. It might be somewhat lower or higher, but it would not equal the population mean exactly. Similarly, if you took a second sample of 40 women from the same population, you would not expect the mean of this second sample to equal the mean of the first sample.

For example, consider a normal population with mean μ and variance σ. Assume we repeatedly take samples of a given size from this population and calculate the arithmetic mean for each

sample. This statistic is then called the sample mean. Each sample has its own average value, and the distribution of these averages is called the “sampling distribution of the sample mean”. This distribution is normal since the underlying population is normal, although sampling distributions may also often be close to normal even when the population distribution is not.

The Expected Value of the Sample Mean:

$$E(\bar{x}) = \mu$$

Standard Error of the Sample Mean:

The standard deviation of the sampling distribution of a statistic is referred to as the standard error of that quantity. For the case where the statistic is the sample mean, and samples are independent, the standard error of the mean is**:**

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

Where **σ** is the population standard deviation and **n** is the size (number of items) in the sample. An important implication of this formula is that the sample size must be increased fourfold (multiplied by 4) to achieve half the measurement error.

If all the sample means were very close to the population mean, then the standard error of the mean would be small. On the other hand, if the sample means varied considerably, then the standard error of the mean would be large. To be specific, assume your sample mean is 125 and you estimated that the standard error of the mean is 5. If you had a normal distribution, then it would be likely that your sample mean would be within 10 units of the population mean since most of a normal distribution is within two standard deviations of the mean.

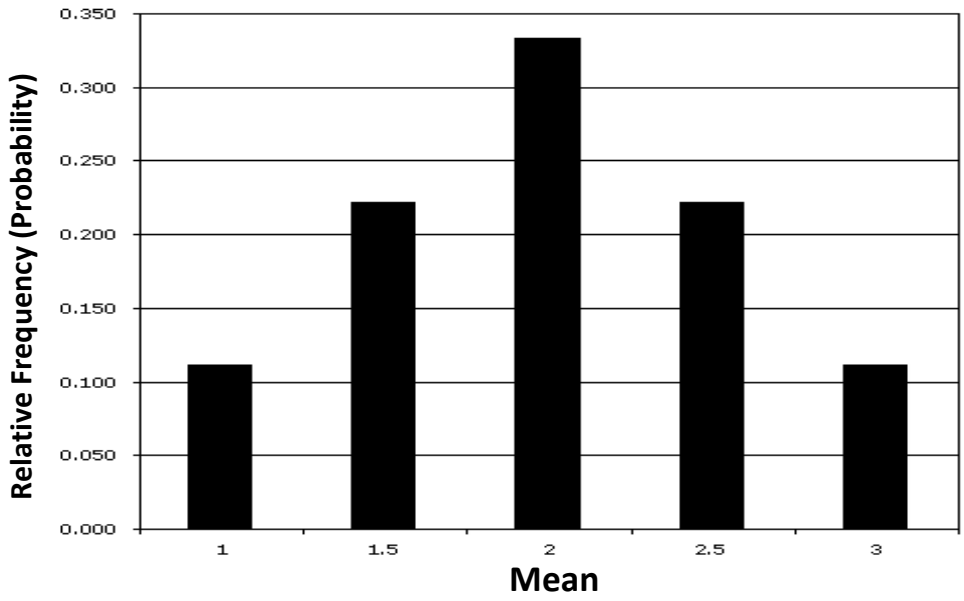
Example (4.1):

A box contains three balls, each with a number on it. Two of the balls are selected randomly (with replacement), and the average of their numbers is computed. The following table shows all the possible outcome of selecting two balls randomly from a population of three.

Sample (Outcome)	Mean	Frequency	Relative Frequency
1,1	1.0	1	0.111
1,2 2,1	1.5	2	0.222
1,3 3,1 2,2	2.0	3	0.333
2,3 3,2	2.5	2	0.222
3,3	3.0	1	0.111

Notice that all the means are either 1.0, 1.5, 2.0, 2.5, or 3.0. The frequencies of these means are shown above. The relative frequencies are equal to the frequencies divided by nine because there are nine possible outcomes.

The figure below shows a relative frequency distribution of the means. This distribution is also a probability distribution since the y-axis is the probability of obtaining a given mean from a sample of two balls in addition to being the relative frequency. The distribution shown in the figure shown below is called the sampling distribution of the mean. Specifically, it is the sampling distribution of the mean for a sample size of 2. For this simple example, the distribution of balls and the sampling distribution are both discrete distributions. The balls have only the numbers 1, 2, and 3, and a sample mean can have one of only five possible values as shown in the above table.



Relative Frequency Distribution of Ball Example

There is an alternative way of conceptualizing a sampling distribution that will be useful for more complex distributions. Imagine that two balls are sampled (with replacement), and the mean of the two balls is computed and recorded. This process is repeated for a second sample, a third sample, and eventually thousands of samples. After thousands of samples are taken and the mean is computed for each, a relative frequency distribution is drawn. The more samples, the closer the relative frequency distribution will come to the sampling distribution shown in the above figure. As the number of samples approaches infinity, the frequency distribution will approach the sampling distribution. This means that you can conceive of a sampling distribution as being a frequency distribution based on a very large number of samples. To be strictly correct, the sampling distribution only equals the frequency distribution exactly when there is an infinite number of samples.

The Classical Central Limit Theorem (CLT):

The central limit theorem formula is being widely used in the probability distribution and sampling techniques. The central limit theorem states that as the sample size gets larger and larger the sample approaches a normal distribution. No matter what the shape of the population distribution is, the fact essentially holds true as the sample size is **over 30**. The central limit theorem essentially has the following characteristics**:**

- Mean of Sample is the same as the mean of the population.
- The standard deviation which is calculated is the same as the standard deviation of the population divided by the square root of the sample size.

A formula for Central Limit Theorem is given by**:**

$$\mathbf{E(\bar{x}) = \mu}$$

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

Where,

$\bar{\mathbf{x}}$ **= Sample Mean**

μ = Population Mean

σ = Population Standard Deviation

$\boldsymbol{\sigma_{\bar{x}}}$ **= Sample Standard Deviation**

n = Sample Size

Therefore, the central limit theorem states that**:**

For large enough sample size, the variable $\mathbf{Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}}$ is normally distributed with a mean of zero and a standard deviation of 1. That is, Z has a standard normal distribution regardless of the shape of the population distribution.

Relevance and Uses of Central Limit Theorem:

- The central limit theorem is widely used in sampling and probability distribution and statistical analysis where a large sample of data is considered and needs to be analyzed in detail.
- The central limit theorem is also used in finance to analyze stocks and index which simplifies many procedures of analysis as generally and most of the times you will have a sample size which is greater than 50.
- Investors of all types rely on the Central Limit Theorem to analyze stock returns, construct portfolios and manage risk.
- A central limit theorem is also used in Binomial probability which places an active role in the analysis of statistical data in detail.

The Importance of the Central Limit Theorem to Statistical Analysis:

Although the normal distribution occurs frequently and describes many populations; it is by no means the only probability distribution in existence. The value of the Central Limit Theorem is found in its conclusion that regardless of the shape of the population distribution (i.e. samples can come from any type of distribution), the sampling distribution of the mean will form a normal distribution. As a result, we can use the normal distribution when we are sampling from populations that we know nothing about and still know the distribution the sample means. Much of statistical inference is based not on the population we are sampling from, but on the distribution of the sample mean or sample proportion.

The Relationship Between the Sample Size and the Standard Error of the Mean:

Question: Why does the sample size play such an important role in reducing the standard error of the mean? What are the implications of increasing the sample size?

Answer: The standard error is the standard deviation of the population you are sampling from divided by the standard deviation of the sample size. So, mathematically as the sample size increases, the standard error naturally decreases. But there is more to this, because the standard error is the standard deviation of the population of sample means. So, as the sample size increases, the sample means are deviating less and less from the true population mean. Hence, as we sample more, we get statistics which are closer to the true parameters and our inference methods will improve. This is true for sampling distributions of mean, proportions, and variances.

Example (4.2):

Suppose that the mean height of college students is 70 inches with a standard deviation of 5 inches. If a random sample of 60 college students is taken, what is the probability that the sample average height for this sample will be more than 71 inches?

Solution:

First check to see if the Central Limit Theorem applies. Since **n > 30**, it does. Next, we need to calculate the standard error. To do that we divide the population standard deviation by the square-root of **n**, which gives

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{5}{\sqrt{60}} = 0.645$$

Next, we calculate a z-score using our z-score formula**:**

$$Z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$$

Plugging in gives us:

$$Z = \frac{71-70}{0.645} = 1.55$$

Finally, we look up our z-score in our z-score table to get a p-value. The table gives us a p-value of,

$$P(z > 1.55) = 1 - P(z < 1.55)$$
$$= 1 - 0.9394 = 0.0606$$

Example (4.3):

In a country located in the middle east region, the recorded weights of the male population are following a normal distribution. The mean and the standard deviations are 70 kg and 16 kg respectively. If a person is eager to find the record of 64 males in the population.

(a) What would mean and the standard deviation of the chosen Sample?

(b) Find the probability that the sample mean is at least 75 kg.

Solution:

Here, $\mu = 70$, $\sigma = 16$, $n = 64$

Since the sample size > 30, the central Limit Theorem is applied.

(a) $E(\bar{x}) = \mu = 70$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{16}{\sqrt{64}} = 2$$

(b)

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$$P(\bar{X} \leq 75) = P\left(\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq \frac{75 - 70}{\frac{16}{\sqrt{64}}}\right)$$

$= P(Z \leq 2.5) = 0.9938$ From the Z- table

Example (4.4):

In examining the invoices issued by a company, an auditor finds that the dollar amount of invoices has a mean of \$1,732 and standard deviation of \$298. Which pair of symmetric numbers around the mean make the statement **P(A $\leq \bar{x} \leq$ B) = 0.853** correct for a random sample of 55 invoices?

Solution:

$\mu = 1732$, $\sigma = 298$, $n = 55$, then $\frac{\sigma}{\sqrt{n}} = 40.182$

$P(a \leq \bar{x} \leq b) = 0.853$ giving

$$P\left(\frac{A - 1732}{40.182} \leq Z \leq \frac{B - 1732}{40.182}\right) = 0.835$$

$$P(Z \leq B) = 1 - \left(\frac{1 - 0.835}{2}\right) = 1 - 0.0735$$

$= 0.9265$

From Z-table, $\frac{B - 1732}{40.182} = 1.45$ which gives B = 1790.26

Since $\frac{A + B}{2} = \mu = 1732 \Rightarrow \frac{A + 1790.26}{2} = 1732$

Then, A = 2(1732) – 1790.26

= 1673.74

Exercises for Section 4.2

(4.1) Random samples of size 225 are drawn from a population with mean 100 and standard deviation 20. Find the mean and standard deviation of the sample mean.

(4.2) A population has mean 5.75 and standard deviation 1.02. Random samples of size 81 are taken.

(a) Find the mean and standard deviation of the sample mean.

(b) How would the answers to Part (a) change if the size of the samples were 64 instead of 81?

(4.3) A normally distributed population has mean 25.6 and standard deviation 3.3.

(a) Find the probability that a single randomly selected element X of the population exceeds 30.

(b) Find the mean and standard deviation of $\bar{\mathbf{X}}$ for samples of size 9.

(c) Find the probability that the mean of a sample of size 9 drawn from this population exceeds 30.

(4.4) A population has mean 72 and standard deviation 6.

(a) Find the mean and standard deviation of $\bar{\mathbf{x}}$ for samples of size 45.

(b) Find the probability that the mean of a sample of size 45 will differ from the population mean 72 by at least 2 units, that is, is either less than 70 or more than 74.

(**Hint:** One way to solve the problem is to first find the probability of the complementary event).

(4.5) Scores on a common final exam in a Statistics Course are normally distributed with mean 72.7 and standard deviation 13.1.

(a) Find the probability that the score X on a randomly selected exam paper is between 70 and 80.

(b) Find the probability that the mean score $\bar{\mathbf{x}}$ of 38 randomly selected exam papers is between 70 and 80.

(4.6) Suppose that in a certain region of the country the mean duration of first marriages that end in divorce is 7.8 years, standard deviation 1.2 years. Find the probability that in a sample of 75 divorces, the mean age of the marriages is at most 8 years.

(4.7) A tire manufacturer states that a certain type of tire has a mean lifetime of 60,000 miles. Suppose lifetimes are normally distributed with standard deviation σ = 3,500 miles.

(a) Find the probability that if you buy one such tire, it will last only 57,000 or fewer miles. If you had this experience, is it particularly strong evidence that the tire is not as good as claimed?

(b) A consumer group buys five such tires and tests them. Find the probability that average lifetime of the five tires will be 57,000 miles or less. If the mean is so low, is that particularly strong evidence that the tire is not as good as claimed?

(4.8) The weights of apples from a large farm are normally distributed with a mean of 380 gm and a standard deviation of 28 gm.

(a) A single apple is selected at random from this farm. What is the probability that it weighs more 400 gm?

(b) Three apples are selected at random from this farm. What is the probability that their mean weight is greater than 400 gm.?

(c) Explain why the probabilities in **Part (a)** and **Part (b)** are not equal.

(4.9) (MCQ Question): According to the central limit theorem, the sampling distribution of the sample mean can be approximated by the normal distribution as the

(a) number of samples gets "large enough"

(b) sample size gets "large enough"

(c) population standard deviation increases

(d) sample standard deviation decreases

4.3 Distribution of Difference in Sample Means:

Statistics problems often involve comparisons between two independent sample means. We are going to explain how to compute probabilities associated with differences between means.

Difference Between Means:

Theory: Suppose we have two populations with means equal to μ_1 and μ_2. Suppose further that we take all possible samples of size n_1 and n_2. And finally, suppose that the following assumptions are valid.

- The size of each population is large relative to the sample drawn from the population. That is, N_1 is large relative to n_1, and N_2 is large relative to n_2. (In this context, populations are considered to be large if they are at least **20** times bigger than their sample), that is $n < 0.05N$.
- The samples are independent; that is, observations in population 1 are not affected by observations in population 2, and vice versa.
- The set of differences between sample means is normally distributed. This will be true if each population is normal or if the sample sizes are large. (Based on the central limit theorem, sample sizes of 40 would probably be large enough).

Given these assumptions, the sampling distribution of the difference between means can be thought of as the distribution that would result if we repeated the following three steps over and over again:

1. Sample n_1 scores from Population 1 and n_2 scores from Population 2;
2. Compute the means of the two samples ($\bar{x}_1$ and $\bar{x}_2$);

3. Compute the difference between means $\bar{x}_1 - \bar{x}_2$. The distribution of the differences between means is the sampling distribution of the difference between means.

Then, the sampling distribution of the difference between two means as follows**:**

The Mean of the Sampling Distribution of Difference in Sample Means:

- The **expected value** of the difference between all possible sample means is equal to the difference between population means. Thus,

$$\mu_{\bar{x}_1} - \mu_{\bar{x}_1} = \mu_1 - \mu_2$$

or

$$\mu_d = E(\bar{x}_1 - \bar{x}_2)$$
$$= E(\bar{x}_1) - E(\bar{x}_2) = \mu_1 - \mu_2$$

The mean of the distribution of differences between sample means is equal to the difference between population means.

- **The standard deviation** of the difference between sample means **(σ_d)** is approximately equal to**:**

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\sigma^2_{\bar{x}_1} + \sigma^2_{\bar{x}_2}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

or

$$\sigma_d = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

The derivation starts with a recognition that the variance of the difference between independent random variables is equal to the sum of the individual variances. Thus,

$$\sigma^2_d = \sigma^2_{(\bar{x}_1 - \bar{x}_2)} = \sigma^2_{\bar{x}_1} + \sigma^2_{\bar{x}_2}$$

If the two population sizes, N_1 and N_2, are both large relative to n_1 and n_2, respectively, then

$$\textbf{Var}(\bar{x}_1) = \frac{\sigma_1^2}{n_1} \quad \text{and} \quad \textbf{Var}(\bar{x}_2) = \frac{\sigma_2^2}{n_2}$$

$$\textbf{Var}(\bar{x}_1 - \bar{x}_2) = \sigma_d^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

That is,

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\sigma_{\bar{x}_1}^2 + \sigma_{\bar{x}_2}^2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

The standard deviation of sampling distribution of differences between two means, i.e. $(\bar{x}_1 - \bar{x}_2)$ is also called as the standard error of $(\bar{x}_1 - \bar{x}_2)$.

From the mean and standard deviation of the difference between sample means, the formula of Z is given by:

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma_d}$$

and $$(\bar{x}_1 - \bar{x}_2) \sim N\{(\mu_1 - \mu_2), \sigma_d\}$$

Where σ_d is given before.

Example (4.5):

For boys, the average number of absences in the first grade is 15 with a standard deviation of 7; for girls, the average number of absences is 10 with a standard deviation of 6.
In a nationwide survey, suppose 100 boys and 50 girls are sampled. What is the probability that the male sample will have at most three more days of absences than the female sample?

Solution:

The solution involves three or four steps, depending on whether you work directly with raw scores or z-scores. The "raw score" solution appears below:

- To find the mean difference (male absences minus female absences) in the population**:**

  $\mu_d = \mu_1 - \mu_2 = 15 - 10 = 5$

- To find the standard deviation of the difference**:**

$$\sigma_d = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{7^2}{100} + \frac{6^2}{50}}$$

$$= \sqrt{\frac{49}{100} + \frac{36}{50}} = \sqrt{0.49 + 0.72} = \sqrt{1.21} = 1.1$$

- To find the probability, this problem requires us to find the probability that the average number of absences in the boy sample minus the average number of absences in the girl sample is less than 3. That is,

$$P\{(\bar{x}_1 - \bar{x}_2) \leq 3\}$$

$$P\{(\bar{x}_1 - \bar{x}_2) \leq 3\} = P\left(Z \leq \frac{3-5}{1.1}\right) = P(Z \leq -1.82)$$

The probability that the difference between sample means will be no more than 3 days is

$$P\{(\bar{x}_1 - \bar{x}_2) \leq 3\} = P(Z \leq -1.82)$$

$$= 1 - P(Z \leq 1.82)$$

$$= 1 - 0.9656 = 0.0344$$

Example (4.6):

Suppose that the starting salaries of the employees in Company (1) are normally distributed with a mean of $62,000 and a standard deviation of $14,500. The starting salaries of the employees in Company (2) are normally distributed with a mean of $60,000 and a standard deviation of $18,300. If a random sample of 50 and a random sample of 60 are selected from Company (1) and Company (2) respectively, what is the

probability that the sample mean starting salary of Company (1) will exceed that of Company (2)?

Solution:

We want to determine $P\{(\bar{x}_1 - \bar{x}_2) > 0\}$. We know that $\bar{X}_1 - \bar{X}_2$ is normally distributed with mean $\mu_1 - \mu_2 = 62{,}000 - 60{,}000 = 2{,}000$ and standard deviation

$$\sigma_d = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$= \sqrt{\frac{(14{,}500)^2}{50} + \frac{(18{,}300)^2}{60}} = 3{,}128$$

We can standardize the variable and refer to Z-table**:**

$$P\{(\bar{x}_1 - \bar{x}_2) > 0\} = P\left\{\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma_d} > \frac{0 - 2{,}000}{3{,}128}\right\}$$

$$= P(Z > -0.64) = 1 - P(Z \leq -0.64)$$

$$= 1 - \{1 - P(Z \leq 0.64)\}$$

$$= P(Z \leq 0.64) = 0.7389$$

Example (4.7):

A professor of statistics noticed that the marks in his course are normally distributed. He has also noticed that his morning class average 37 with a standard deviation of 12 on their final exam. His afternoon class average 77 with a standard deviation of 10. What is the probability that the mean mark of four randomly selected students from a morning class is at most two marks less than the average mark of four randomly selected students from an afternoon class?

Solution:

We want to determine $P\{(\bar{x}_1 - \bar{x}_2) \leq 2\}$. We know that $\bar{X}_1 - \bar{X}_2$ is normally distributed with mean

$\mu_1 - \mu_2 = 73 - 77 = -4$

and standard deviation

$$\sigma_d = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{(12)^2}{4} + \frac{(10)^2}{4}}$$

$$= 7.81$$

We can standardize the variable and refer to Z-table**:**

$$P\{(\bar{x}_1 - \bar{x}_2) \leq 2\} = P\left\{\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma_d} > \frac{2 - (73 - 77)}{7.81}\right\}$$

$$= P(Z \leq 0.77) = 1 - 0.7794$$

$$= 0.2206$$

Exercises for Section 4.3

(4.10) Population 1 has a mean of 20 and a variance of 100. Population 2 has a mean of 15 and a variance of 64. You sample 20 scores from Population 1 and 16 scores from Population 2.

(a) What is the mean of the sampling distribution of the difference between means (Population 1 - Population 2)?

(b) What is the standard deviation of the sampling distribution of the difference between means (Pop1 - Pop 2)?

(4.11) Suppose we have two normal populations with mean and standard deviations listed below. If random samples of size 25 are drawn from each population, what is the probability the mean of sample (1) is greater than the mean of sample (2)?

Population (1): $\mu = 40$, $\sigma = 6$

Population (2): $\mu = 38$, $\sigma = 8$

(4.12) Every day, thousands of people at an airport pass through security on one of two levels: level A or level B. Suppose that, on average, it takes people 26 minutes to pass through security on level A with a standard deviation of 7.5 minutes. On level B, the mean and standard deviation are 24 and 4 minutes, respectively. Each day, the airport looks at separate random samples of 100 people from each level.

(a) What are the mean and standard deviation (in minutes) of the sampling distribution of $\mathbf{(\bar{x}_A - \bar{x}_B)}$.

(b) What do we know about the shape of the sampling distribution of $\mathbf{(\bar{x}_A - \bar{x}_B)}$?

(c) On any given day, what is the probability that the sample mean from level A is 4 minutes higher than the sample mean from level B?

(d) On any given day, what is the approximate probability that the mean from level B exceeds the mean from level A?

4.4 Distribution of Sample Proportion:

The sampling distribution of the sample proportion If repeated random samples of a given size n are taken from a population of values for a categorical variable, where the proportion in the category of interest is p, then the mean of all sample proportions $\widehat{\mathbf{P}}$ **(P- hat)** is the population proportion **(P)**, that is

$$\mathbf{E(\widehat{P}) = P}$$

As for the spread of all sample proportions, theory dictates the behavior much more precisely than saying that there is less spread for larger samples. In fact, the standard deviation of all sample proportions is directly related to the sample size, n as indicated below.

The Standard deviation of all sample properties ($\widehat{\mathbf{P}}$) is exactly:

$$\mathbf{S_E\,(\widehat{P}) = \sqrt{\frac{P(1-P)}{n}}}$$

Since the sample size **n** appears in the denominator of the square root, the standard deviation does decrease as sample size increases. Finally, the shape of the distribution of $\widehat{\mathbf{P}}$ will be approximately normal as long as the sample size **n** is large enough. The convention is to require both **np** and **np(1 – p)** to be at least 10.

From the mean and standard deviation of the sample proportion $\widehat{\mathbf{P}}$, the formula of Z is given by**:**

$$\mathbf{Z = \frac{\widehat{P} - P}{\sqrt{\frac{P(1-P)}{n}}}}$$

This means

$$\frac{\widehat{\mathbf{P}}-\mathbf{P}}{\sqrt{\frac{\mathbf{P}(\mathbf{1}-\mathbf{P})}{\mathbf{n}}}} \sim \mathbf{N(0\ ,\ 1)}$$

Example (4.8):

A random sample of 100 students is taken from the population of all students in a university, for which the overall proportion of females is 0.6.

(a) What is the probability that sample proportion **($\widehat{\mathbf{P}}$)** is less than or equal to 0.58?

(b) There is a 95% chance that the sample proportion **($\widehat{\mathbf{P}}$)** falls between what two values?

Solution:

In our example, n = 100(sample size) and P = 0.6.

Note that nP = 60 > 10 and n(1 – p) = 40 > 10, therefore we can conclude that $\widehat{\mathbf{P}}$ is approximately a normal distribution with mean P = 0.6 and standard deviation $\sqrt{\frac{0.6(1-0.6)}{100}}$

(a) To find $P(\widehat{P} \leq 0.58)$

First note that the distribution of $(\widehat{P})$ has mean P = 0.6 and standard deviation = $\sqrt{\frac{P(1-P)}{n}} = \sqrt{\frac{0.6(1-0.6)}{100}} = 0.05$

we standardize 0.58 into a z-score by subtracting the mean and dividing the result by the standard deviation.

$$Z = \frac{0.58 - 0.6}{0.05} = -0.4$$

Then we can find the probability as follows**:**

$$P(\hat{P} \leq 0.58) = P(Z \leq \frac{0.58 - 0.6}{0.05}) = P(Z \leq -0.4)$$

$$= 1 - P(Z \leq 0.4) = 1 - 0.6554 = 0.3446$$

(b) The Standard Deviation Rule applies**:** the probability is approximately 0.95 that $\hat{P}$ falls within 2 standard deviations of the mean, that is, between 0.6 – 2(0.05) and 0.6 + 2(0.05). There is roughly a 95% chance that $\hat{P}$ falls in the interval [0.5 , 0.7] for samples of this size.

Example (4.9):

A record store owner finds that 20% of the customers entering her store make a purchase. One morning 180 people, who can be regarded as a random sample of all customers, enter the store.

(a) What is the mean of the distribution of the sample proportion of customers making a purchase?

(b) What is the standard deviation of the sample proportion?

(c) What is the probability that the sample proportion is less than 0.15?

Solution:

In this example, n = 400 (sample size) and P = 0.2.
Note that nP = 80 > 10 and n(1 – p) = 329 > 10, therefore we can conclude that $\hat{P}$ is approximately a normal distribution.

(a) $E(\hat{P}) = P = 0.2$

(b) $Var(\hat{P}) = \sqrt{\frac{P(1-P)}{n}} = \sqrt{\frac{0.2(1-0.2)}{400}} = 0.02$

(c) $P(\hat{P} \leq 0.25) = P(Z \leq \frac{0.25-0.2}{0.02}) = P(Z \leq 2.5)$

$= 0.9938$

Exercises for Section 4.4

(4.13) The proportion of a population with a characteristic of interest is **P = 0.82**. Find the mean and standard deviation of the sample proportion $\widehat{\mathbf{P}}$ obtained from random samples of size 900.

(4.14) Samples of size **n** produced sample proportions $\widehat{\mathbf{P}}$ as shown. In each case decide whether or not the sample size is large enough to assume that the sample proportion $\widehat{\mathbf{P}}$ is normally distributed.

(a) n = 30, $\widehat{\mathbf{P}}$ = 0.72 **(b)** n = 40, $\widehat{\mathbf{P}}$ = 0.12

(c) n = 75, $\widehat{\mathbf{P}}$ = 0.84

(4.15) A random sample of size 121 is taken from a population in which the proportion with the characteristic of interest is **P = 0.47**. Find the following probabilities**:**

(a) $P(0.45 \le \widehat{\mathbf{P}} \le 0.50)$ **(b)** $P(\widehat{\mathbf{P}} > 0.50)$

(4.16) A random sample of size 225 is taken from a population in which the proportion with the characteristic of interest is P = 0.34. Find the indicated probabilities.

(a) $P(0.25 \le \widehat{\mathbf{P}} \le 0.40)$ **(b)** $P(\widehat{\mathbf{P}} \ge 0.35)$

(4.17) A random sample of size 900 is taken from a population in which the proportion with the characteristic of interest is p = 0.62. Find the following probabilities**:**

(a) $P(0.60 \le \widehat{\mathbf{P}} \le 0.64)$ **(b)** $P(0.57 \le \widehat{\mathbf{P}} \le 0.67)$

(4.18) Suppose that 8% of all males suffer from some form of color blindness. Find the probability that in a random sample of 250 men at least 10% will suffer from some form of color blindness. First verify that the sample is sufficiently large to use the normal distribution.

(4.19) Suppose that 2% of all cell phone connections by a certain provider are dropped. Find the probability that in a random sample of 1,500 calls at most 40 will be dropped. First verify that the sample is sufficiently large to use the normal distribution.

(4.20) An outside financial auditor has observed that about 4% of all documents he examines contain an error of some sort. Assuming this proportion to be accurate, find the probability that a random sample of 700 documents will contain at least 30 with some sort of error. You may assume that the normal distribution applies.

(4.21) Suppose 7% of all households have no home telephone but depend completely on cell phones. Find the probability that in a random sample of 450 households, between 25 and 35 will have no home telephone. You may assume that the normal distribution applies.

(4.22) An ordinary die is "fair" or "balanced" if each face has an equal chance of landing on top when the die is rolled. Thus, the proportion of times a three is observed in a large number of tosses is expected to be close to 1/6 or 0.1667. Suppose a die is rolled 240 times.

- **a.** Find the probability that a fair die would produce a proportion of 0.15 or less. You may assume that the normal distribution applies.
- **b.** Give an interpretation of the result in **Part (a)**. How strong is the evidence that the die is not fair?
- **c.** Suppose the sample proportion 0.15came from rolling the die 400 times instead of only 225 times. Rework **Part (a)** under these circumstances.
- **d.** Give an interpretation of the result in **Part (c)**. How strong is the evidence that the die is not fair?

4.5 Distribution of Difference in Sample Proportions:

Statistics problems often involve comparisons between two independent sample proportions. This section explains how to compute probabilities associated with differences between proportions.

Difference Between Proportions:

Theory: Suppose we have two populations with proportions equal to $\mathbf{P_1}$ and $\mathbf{P_2}$. Suppose further that we take all possible samples of size $\mathbf{n_1}$ and $\mathbf{n_2}$. And finally, suppose that the following assumptions are valid.

- The size of each population is large relative to the sample drawn from the population. That is, $\mathbf{N_1}$ is large relative to $\mathbf{n_1}$, and $\mathbf{N_2}$ is large relative to $\mathbf{n_2}$. (In this context, populations are considered to be large if they are at least 20 times bigger than their sample, i.e. **n < 0.05 N**).

- The samples from each population are big enough to justify using a normal distribution to model differences between proportions. The sample sizes will be big enough when the following conditions are met:
  $\mathbf{n_1P_1 \geq 10}$, $\mathbf{n_1(1-P_1) \geq 10}$, $\mathbf{n_2P_2 \geq 10}$, and $\mathbf{n_2(1-P_2) \geq 10}$ (This criterion requires that at least 40 observations be sampled from each population. When P_1 or P_2 is more extreme than 0.5, even more observations are required).

- The samples are independent; that is, observations in population 1 are not affected by observations in population 2, and vice versa.

Given these assumptions, we know the following**:**

- The set of differences between sample proportions will be normally distributed. We know this from the central limit theorem.

- The expected value of the difference between all possible sample proportions is equal to the difference between population proportions. Thus,

$$E(\hat{P}_1 - \hat{P}_2) = P_1 - P_2$$

- The standard deviation of the difference between sample proportions (σ_d) is approximately equal to:

$$\sigma_d = \sqrt{\left[\frac{P_1(1-P_1)}{n_1}\right] + \left[\frac{P_2(1-P_2)}{n_2}\right]}$$

The derivation starts with a recognition that the variance of the difference between independent random variables is equal to the sum of the individual variances. Thus,

$$\sigma_d^2 = \sigma_{(P_1 - P_2)}^2 = \sigma_1^2 + \sigma_2^2$$

If the populations N_1 and N_2 are both large relative to n_1 and n_2, respectively, then

$$\sigma_1^2 = \frac{P_1(1-P_1)}{n_1} \quad \text{and} \quad \sigma_2^2 = \frac{P_2(1-P_2)}{n_2}$$

Therefore,

$$\sigma_d^2 = \left[\frac{P_1(1-P_1)}{n_1}\right] + \left[\frac{P_2(1-P_2)}{n_2}\right]$$

and

$$\sigma_d = \sqrt{\left[\frac{P_1(1-P_1)}{n_1}\right] + \left[\frac{P_2(1-P_2)}{n_2}\right]}$$

From the mean and standard deviation of the difference between sample proportions $(\hat{P}_1 - \hat{P}_2)$, the formula of **Z** is given by:

$$Z = \frac{(\hat{P}_1 - \hat{P}_2) - (P_1 - P_2)}{\sigma_d}$$

Where σ_d is given above.

This means that

$$\frac{\left(\hat{P}_1 - \hat{P}_2\right) - (P_1 - P_2)}{\sqrt{\left[\frac{P_1(1-P_1)}{n_1}\right] + \left[\frac{P_2(1-P_2)}{n_2}\right]}} \sim N(0, 1)$$

Comment: Let's look at the relationship between the sampling distribution of differences between sample proportions and the sampling distributions for the individual sample proportions given in Section 4.4 We compare these distributions in the following table:

Sampling Distribution	Sample Proportions from Population 1	Sample Proportions from Population 2	All Differences in Sample Proportions from the two Populations
Mean	P_1	P_2	$P_1 - P_2$
Standard Error	$\sqrt{\frac{P_1(1-P_1)}{n_1}}$	$\sqrt{\frac{P_2(1-P_2)}{n_2}}$	$\sqrt{\frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}}$
Conditions for Use of a Normal Distribution	• n_1p_1 and $n_1(1-p_1)$ are greater than 10 • Expected successes and failures are at least 10	• n_2p_2 and $n_2(1-p_2)$ are greater than 10 • Expected successes and failures are at least 10	• n_1p_1 and $n_1(1-p_1)$ are greater than 10 • n_2p_2 and $n_2(1-p_2)$ are greater than10 • Expected successes and failures in both samples at least 10

Notice the relationship between the means:

- The mean of the differences is the difference of the means. This makes sense. The mean of each sampling distribution of individual proportions is the population proportion, so the mean of the sampling distribution of differences is the difference in population proportions.

Notice the relationship between standard errors:

- The standard error of differences relates to the standard errors of the sampling distributions for individual proportions. Look at the terms under the square roots. Since we add these terms, the standard error of differences is always larger than the standard error in the sampling distributions of individual proportions. In other words, there is more variability in the differences.

Now, we work through some examples to show how to apply the theory presented before.

Example (4.10):

In one state, 52% of the voters are Republicans, and 48% are Democrats. In a second state, 47% of the voters are Republicans, and 53% are Democrats. Suppose 100 voters are surveyed from each state. Assume the survey uses simple random sampling.
What is the probability that the survey will show a greater percentage of Republican voters in the second state than in the first state?

Solution:

We will solve this example following the detailed steps:
Let $\mathbf{P_1}$ = the proportion of Republican voters in the first state,
$\mathbf{P_2}$ = the proportion of Republican voters in the second state,
$\widehat{\mathbf{P}}_1$ = the proportion of Republican voters in the sample from the first state, and
$\widehat{\mathbf{P}}_2$ = the proportion of Republican voters in the sample from the second state.

The number of voters sampled from the first state $\mathbf{n_1 = 100}$, and the number of voters sampled from the second state: $\mathbf{n_2 = 100}$.

The solution involves **four** steps.

- Make sure the samples from each population are big enough to model differences with a normal distribution. Because $n_1P_1 = 100 \times 0.52 = 52$,
  $n_1(1 - P_1) = 100 \times 0.48 = 48$,
  $n_2P_2 = 100 \times 0.47 = 47$,
  and $n_2(1 - P_2) = 100 \times 0.53 = 53$ are each greater than 10, the sample size is large enough.

- Find the mean of the difference in sample proportions**:**

$$E(\hat{P}_1 - \hat{P}_2) = P_1 - P_2 = 0.52 - 0.47 = 0.05$$

- Find the standard deviation of the difference.

$$\sigma_d = \sqrt{\left[\frac{P_1(1-P_1)}{n_1}\right] + \left[\frac{P_2(1-P_2)}{n_2}\right]}$$

$$= \sqrt{\frac{(0.52)(0.48)}{100} + \frac{(0.47)(0.53)}{100}}$$

$$= \sqrt{0.002496 + 0.002491}$$

$$= \sqrt{0.004987} = 0.0706$$

- Find the probability. This problem requires us to find the probability that $\hat{\mathbf{P}}_1$ is less than $\hat{\mathbf{P}}_2$. This is equivalent to finding the probability that $\hat{\mathbf{P}}_1 - \hat{\mathbf{P}}_2$ is less than zero. To find this probability, we need to transform the random variable $(\hat{\mathbf{P}}_1 - \hat{\mathbf{P}}_2)$ into a z-score. That transformation appears below.

$$Z = \frac{(\hat{P}_1 - \hat{P}_2) - (P_1 - P_2)}{\sigma_d}$$

$$= \frac{0 - 0{,}05}{0{,}0706} = -0.71$$

$$P\{(\hat{P}_1 - \hat{P}_2) \leq 0\} = P(Z \leq -0.71) = 1 - P(Z \leq 0.71)$$

Using **Z-table**, we find that the probability of a z-score being 0.71 or less is 0.7611.

Therefore, the probability that the survey will show a greater percentage of Republican voters in the second state than in the first state is (1 - 0.76) = 0.24

Example (4.11):

Two drugs, cure rates of 60% and 65%, what is probability that drug 1 will cure more in the sample than drug 2 if we sample 200 taking each drug?

Solution:

Let $\mathbf{P_1}$ = the cure proportion of the first drug.

$\mathbf{P_2}$ = the cure proportion of the second drug,

$\hat{\mathbf{P}}_\mathbf{1}$ = the cure proportion in the sample taking the first drug.

$\hat{\mathbf{P}}_\mathbf{2}$ = the cure proportion in the sample taking the second drug.

The number of patients sampled for the first drug $\mathbf{(n_1) = 200}$, and the number of voters sampled from the second state $\mathbf{(n_2) = 200}$.

The solution involves **four** steps:

- Make sure the samples from each population are big enough to model differences with a normal distribution. Because $n_1P_1 = 200 \times 0.6 = 120$,
  $n_1(1 - P_1) = 200 \times 0.4 = 80$, $n_2P_2 = 200 \times 0.65 = 130$,
  and $n_2(1 - P_2) = 200 \times 0.35 = 70$ are each greater than 10, the sample size is large enough.

- Find the mean of the difference in sample proportions**:**

  $E(\hat{P}_1 - \hat{P}_2) = P_1 - P_2 = 0.6 - 0.65 = -0.05$

- Find the standard deviation of the difference.

$$\sigma_d = \sqrt{\left[\frac{P_1(1-P_1)}{n_1}\right] + \left[\frac{P_2(1-P_2)}{n_2}\right]}$$

$$= \sqrt{\frac{(0.6)(0.4)}{200} + \frac{(0.65)(0.35)}{200}}$$

$$= \sqrt{0.0012 + 0.0011} = \sqrt{0.0023} = 0.048$$

- Find the probability. This problem requires us to find the probability that $\widehat{P}_1$ is less than $\widehat{P}_2$. This is equivalent to finding the probability that $\widehat{P}_1 - \widehat{P}_2$ is less than zero. To find this probability, we need to transform the random variable ($\widehat{P}_1 - \widehat{P}_2$) into a z-score. That transformation appears below.

$$Z = \frac{(\widehat{P}_1 - \widehat{P}_2) - (P_1 - P_2)}{\sigma_d}$$

$$= \frac{0 - (-0{,}05)}{0{,}048} = 1.04$$

$$P\{(\widehat{P}_1 - \widehat{P}_2) > 0\} = 1 - P\{(\widehat{P}_1 - \widehat{P}_2) \leq 0\}$$
$$= 1 - P(Z \leq 1.04)$$

Using **Z-table**, we find that the probability of a z-score being 1.04 or less is 0.8508.

Therefore, the probability that the survey will show a greater percentage of Republican voters in the second state than in the first state is (1 – 0.8508) = 0.1492

Exercises for Section 4.5

(4.23) A candidate is running for office, and she wants to know how much support she has in two different districts. She doesn't know it, but 45% of the 8000 voters in District A support her, while 40% of the 6500 voters in District B support her. She hires a polling firm to take separate random samples of 100 voters from each district. The firm will then look at the difference between the proportions of voters who support her in each sample ($\widehat{\mathbf{P}}_A - \widehat{\mathbf{P}}_B$).

(a) What are the mean and standard deviation of the sampling distribution of $\mathbf{(\widehat{P}_A - \widehat{P}_B)}$?

(b) What will be the shape of the sampling distribution of $\mathbf{(\widehat{P}_A - \widehat{P}_B)}$ and why?

(c) What is the probability that the sample from district B shows more support for her than the sample from district A?

(4.24) A company has two offices, one in City (A), and the other in City (B).

- 85% of the employees at City (A) office are younger than 40 years old.
- 81% of the employees at City (B) office are younger than 40 years old.

The company plans on taking separate random samples of 50 employees from each office. They'll look at the difference between the proportions of employees in each sample that are younger than 40 years old $\mathbf{(\widehat{P}_A - \widehat{P}_B)}$.

What is the probability that the difference between the two samples is greater than 10 percentage points? That is greater than 0.1; $\mathbf{P\{(\widehat{P}_A - \widehat{P}_B)\} > 0.1}$.

Chapter (5)
Estimation of Population Parameters

Contents

Chapter 5
Estimation of Population Parameters

5.1 Estimation procedures for One Population

5.1.1 Estimation of a Population Mean:

The most fundamental point and **interval estimation** process involves the estimation of a population mean. Suppose it is of interest to estimate the population mean, **μ**, for a quantitative variable. Data collected from a simple random sample can be used to compute the **sample mean**, $\bar{x}$, where the value of $\bar{x}$ provides a **point estimate** of **μ**.

When the **sample mean** is used as a **point estimate** of the population mean, **μ**, some error can be expected owing to the fact that a sample, or subset of the population, is used to compute the point estimate. The absolute value of the difference between the sample mean, $\bar{x}$, and the population mean, **μ**, written $|\bar{x} - \mu|$, is called the **sampling error**. Interval estimation incorporates a probability statement about the magnitude of the sampling error. The **sampling distribution** of $\bar{x}$, discussed before, provides the basis for such a statement.

Statisticians have shown that the mean of the sampling distribution of $\bar{x}$ is equal to the population mean, **μ**, and that the **standard deviation** is given by $\frac{\sigma}{\sqrt{n}}$, where **σ** is the **population standard deviation** which is usually unknown. The standard deviation of a sampling distribution in this case is called the **standard error (S_E)**, that is

$$S_E = \frac{S}{\sqrt{n}}$$

- **For Large Samples**

For large sample sizes, the **central limit theorem** indicates that the sampling distribution of $\bar{x}$ can be approximated by a normal probability distribution. As a matter of practice, statisticians usually consider samples of size **30** or more to be **large**.

In the large - sample case, a **95%** confidence interval estimate for the population mean is given by

$$\textbf{95\% CL for } \mu \textbf{ is: } \bar{x} \pm 1.96\left(\frac{\sigma}{\sqrt{n}}\right)$$

When the population standard deviation, **σ**, is **unknown**, the sample standard deviation is used to estimate σ in the confidence interval formula, i.e.,

$$\textbf{95\% CL for } \mu \textbf{ is: } \bar{x} \pm 1.96\left(\frac{S}{\sqrt{n}}\right)$$

The quantity $\mathbf{1.96}\left(\frac{\sigma}{\sqrt{n}}\right)$ is often called the **margin of error** for the estimate (It is also called **maximum error** and is denoted by **E**). The quantity $\frac{\sigma}{\sqrt{n}}$ is the **standard error**, and **1.96** is the number of standard errors from the mean necessary to include **95%** of the values in a **normal distribution.** There, we can say that the confidence interval (CL) for the population mean **μ** can be written as:

Point Estimate ± Margin of Error

The interpretation of a 95% confidence interval is that 95% of the intervals constructed in this manner will contain the population mean. Thus, any interval computed in this manner has a 95% confidence of containing the population mean.

By changing the **constant** from 1.96 to **2.58**, a **99%** confidence interval can be obtained. It should be noted from the formula for an interval estimate that a 95% confidence interval is narrower than a 99% confidence interval and as such has a slightly smaller confidence of including the population mean. Lower levels of confidence lead to even more narrow intervals. In practice, a 95% confidence interval is the most widely used.

Owing to the presence of the $\sqrt{\mathbf{n}}$ term in the formula for an interval estimate, the sample size affects the margin of error. Larger sample sizes lead to smaller margins of error. This observation forms the basis for procedures used to select the sample size. Sample sizes can be chosen such that the confidence interval satisfies any desired requirements about the size of the margin of error. This point us to be considered later. According to the central limit theory discussed before, we can say that regardless of the distribution shape of the population, the sampling distribution of $\bar{\mathbf{x}}$ becomes approximately normal as the sample size increases (conservatively $\mathbf{n} \geq \mathbf{30}$).

Now, let us introduce some terms regarding the confidence interval (CI) for a population parameter with the aid of the following phrase: If we say **'we are 95% confident that the mean Statistics score for a group of students is between 70 and 75'**. Then, in this case, **95%** is called **'confidence level'** and is denoted by **100 (1 - α)%** and **70 to 75** is the **95%** confidence interval (CI) for the population mean **μ** and is denoted by **CI**.

70 is the **lower limit** of the **95% CI** for the population mean **μ** and is denoted by **LL**.

75 is the **upper limit** of the **95% CI** for the population mean **μ** and is denoted by **UL**.

75 – 70 = 5 is the width of the **95% CI** for the population mean **μ** and is denoted by **W**.

In general, the confidence interval of the population mean **μ** is given by**:**

$$\textbf{Point Estimate} \pm Z_{\frac{\alpha}{2}}\textbf{(Standard Error)}$$

where

Point estimate for **μ** is $\bar{x}$,

$Z_{\frac{\alpha}{2}}$ **= 1.96** for a **95% CI ;** $Z_{\frac{\alpha}{2}} = Z_{\frac{0.05}{2}} = Z_{0.025}$, and

2.58 for a **99% CI** ; $Z_{\frac{\alpha}{2}} = Z_{\frac{0.01}{2}} = Z_{0.005}$,

Standard Error = $\left(\frac{\sigma}{\sqrt{n}}\right)$ **or** $\left(\frac{S}{\sqrt{n}}\right)$ as the case may be, and

$Z_{\frac{\alpha}{2}}\left(\frac{S}{\sqrt{n}}\right)$ = **Margin of error = E =** $\frac{1}{2}$**(Width)**

Hint: Point Estimate = $\frac{\textbf{Lower Limit (LL)} + \textbf{Upper Limit(UL)}}{2}$

Example 5.1

Suppose a random sample of size n = 100 has been selected and the sample mean is found to be $\bar{x}$ = 67.

The population standard deviation is assumed to be σ = 12.

Answer the following questions.

(a) What is the standard error of the mean $\sigma_{\bar{x}}$?

(b) What is the margin of error for a 95% confidence level for **μ**?

(c) What is the 95% confidence interval estimate of μ?

Solution:

(a) The standard error of the mean $\boldsymbol{\sigma_{\bar{x}}}$ is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{12}{\sqrt{100}} = 1.2$$

(b) If (1 – α) = 0.95, then

$Z_{\frac{\alpha}{2}} = Z_{\frac{0.05}{2}} = Z_{0.025} = 1.96$

Therefore, $Z_{\frac{\alpha}{2}}\sigma_{\bar{x}} = (1.96)(1.2) = 2.352$

(c) The **95% confidence interval** estimate of **μ** is

$\bar{x} \pm Z_{\frac{\alpha}{2}}\sigma_{\bar{x}} = 67 \pm 2.352$

or [64.648, 69.352]

Example 5.2

If a **99%** confidence interval is **[228 , 232]** for a population with **σ = 10**, what is **n**?

Solution:

Margin of error = (232 – 228) / 2 = 2

$$= Z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right) = 2.58\left(\frac{10}{\sqrt{n}}\right)$$

$$\sqrt{n} = \frac{(2.58)(10)}{2}$$

giving $n = \left[\frac{(2.58)(10)}{2}\right]^2 \approx 166$

Example 5.3

In a study of the amount STAT students are spending each term on text- books, data were collected on n = 81 students. In previous studies, the population standard deviation has been σ = $24.

(a) What is the margin of error at the 99% confidence level?

(b) If the mean from the most recent sample was $\bar{x}$ = $288, what is the 99% confidence interval estimate of the population mean μ?

Solution:

(a) The margin of error is given by

If (1 – α) = 0.99 then

$Z_{\frac{\alpha}{2}} = Z_{\frac{0.01}{2}} = Z_{0.005} = 2.58$

$Z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right) = 2.58\left(\frac{24}{\sqrt{81}}\right) = 6.87$

(b) The 99% CI is as follows

The **99% confidence interval** estimate of **μ** is

$\bar{x} \pm Z_{\frac{\alpha}{2}}\sigma_{\bar{x}} = 288 \pm 6.87$ or [281.13 , 294.87]

Example 5.4

A survey finds that the average British household is expected to spend £900 on holiday-related expenses during the next Christmas period. This amount covers not only gifts, but food, beverages, and decorations. Assume the study is based on n = 100 randomly sampled households throughout the U.K. Assume that the sample standard deviation is s = £200.

(a) What is the 95% confidence interval estimate of the population mean amount-to-be-spent μ during the 2022 Christmas holiday period?

(b) What is the 99% confidence interval estimate of the population mean amount – to – be – spent (μ) during the next Christmas holiday period?

Solution:

(a) For a 95% CI of μ, $Z_{\frac{\alpha}{2}} = Z_{\frac{0.05}{2}} = Z_{0.025} = 1.96$

The **95% confidence interval** estimate of **μ** is

$\bar{x} \pm Z_{\frac{\alpha}{2}}S_{\bar{x}} = 900 \pm 1.96\left(\frac{200}{\sqrt{100}}\right)$

$= 900 \pm 39.2$ or [860.8, 939.2]

(b) For a **99% CI of μ**, $Z_{\frac{\alpha}{2}} = Z_{\frac{0.01}{2}} = Z_{0.005} = 2.58$

The **99% confidence interval** estimate of **μ** is

$$\bar{x} \pm Z_{\frac{\alpha}{2}} S_{\bar{x}} = 900 \pm 2.58\left(\frac{200}{\sqrt{100}}\right)$$

$$= 900 \pm 51.6 \text{ or } [848.4, 951.6]$$

- **For Small Samples**

The procedure just described for developing interval estimates of a population mean is based on the use of a large sample. In the small-sample case, i.e., where the sample size n is less than **30**, the t distribution is used when specifying the margin of error and constructing a confidence interval estimate. For example, at a 95% level of confidence, a value from the t distribution, determined by the value of n, would replace the 1.96 value obtained from the normal distribution. The t values will always be larger, leading to wider confidence intervals, but, as the sample size becomes larger, the t values get closer to the corresponding values from a normal distribution.

For example, with a sample size of **25**, the t value used would be **2.064**, as compared with the normal probability distribution value of **1.96** in the large - sample case. Therefore, for a small sample of size **25** and using the t- table, the confidence intervals for **μ** are given as

$$\textbf{95\% CI for } \mu \textbf{ is: } \bar{x} \pm 2.064\left(\frac{S}{\sqrt{n}}\right)$$

$$\textbf{99\% CI for } \mu \textbf{ is: } \bar{x} \pm 2.797\left(\frac{S}{\sqrt{n}}\right)$$

In general, the confidence interval of the population mean **μ** for small samples is given by**:**

$$\textbf{Point Estimate} \pm t_{\frac{\alpha}{2}} \textbf{ (Standard Error)}$$

Where, **Point estimate** for **μ** is $\bar{x}$

$t_{\alpha/2,24}$ = Its value depends on the sample size and confidence level. For example

$t_{0.025,24}$ = **2.064** for a **95% CI, and n = 25**

$t_{0.005,24}$ = **2.797** for a **99% CI**, **and n = 25**

Standard Error = $\left(\frac{S}{\sqrt{n}}\right)$ or $\left(\frac{\sigma}{\sqrt{n}}\right)$ as the case may be.

$$\textbf{Margin of error} = t_{\frac{\alpha}{2}}\left(\frac{S}{\sqrt{n}}\right) = \frac{1}{2}\textbf{(Width)}$$

So, the t- distribution is used for estimating the population mean μ if the following conditions are satisfied:

- The population must be **normally distributed**.
- The sample is considered small, i.e. when **n < 30**.
- The **population standard deviation** is **unknown**.

Example 5.5

A sample of size 15 drawn from a normally distributed population with a sample mean of 35 and a sample standard deviation of 14. Construct a 95% confidence interval for the population mean, and interpret its meaning.

Solution:

Since the population is normally distributed, the sample is small, and the population standard deviation is unknown, the formula that applies is

$$\textbf{Point Estimate} \pm t_{\frac{\alpha}{2}}\ \textbf{(Standard Error)}$$

Confidence level 95% means that $\alpha = 1 - 0.95 = 0.05$, so $\frac{\alpha}{2}$ is equal to 0.025. Since the sample size is $n = 15$, there are $n - 1 = 14$

degrees of freedom, one may be 95% confident that the true population in the interval:

95% CI for μ is: $\bar{x} \pm t_{0.025,14}\left(\frac{S}{\sqrt{n}}\right)$

$35 \pm 2.145\left(\frac{14}{\sqrt{15}}\right) = 35 \pm 7.8$ or [27.2 , 42.8]

So, we can say that we are 95% confident that the true population mean is between 27.2 and 42.8.

Example 5.6

A random sample of 12 students drawn from a large university yields mean GPA of 2.7 with sample standard deviation 0.51. Construct a 99% confidence interval for the mean GPA of all students at the university. Assume that the numerical population of GPAs from which the sample is taken has a normal distribution.

Solution:

Since the population is normally distributed, the sample is small, and the population standard deviation is unknown, the t-distribution is applied.

99% CL means that $\alpha = 1 - 0.99 = 0.01$, so $\frac{\alpha}{2} = 0.005$.

Since the Sample size is $n = 12$, there are $n - 1 = 11$ degrees of freedom, one may be 99% confident that the true average GPA of all students at the university is contained in the interval

99% CI for μ is: $\bar{x} \pm t_{0.005,11}\left(\frac{S}{\sqrt{n}}\right)$

$2.7 \pm 3.106\left(\frac{0.51}{\sqrt{12}}\right) = 2.7 \pm 0.5$ or [2.2 , 3.2]

Example 5.7

For **Example 5.3**, find the 99% confidence interval for the mean stat score of all students using the sample size 225 instead of 81. Compare between the two results.

Solution:

For a **99% CI of** μ, $Z_{\frac{\alpha}{2}} = Z_{\frac{0.01}{2}} = Z_{0.005} = 2.58$

The **99% confidence interval** estimate of **μ** is

$$\bar{x} \pm Z_{\frac{\alpha}{2}}\sigma_{\bar{x}} = 288 \pm 2.58\left(\frac{24}{\sqrt{225}}\right)$$

$$= 288 \pm 4.13 \quad \text{or} \quad [283.87 , 292.13]$$

The summary statistics in the two samples are the same, but the 90%confidence interval for the average Stat scores of all students at the university in "Example 5.6" is shorter than the 99% confidence interval in "Example 5.3". This is partly because in "Example 5.6" the sample size is larger; there is more information pertaining to the true value of μ in the larger data set than in the small one.

Hint: The larger the sample size, the shorter the width of the confidence interval, and thus the greater the accuracy of the estimation.

5.1.2 Estimation of a Population Proportion: (Large Samples)

Consider the percentage of people in favor of a four-day work week, the percentage of men who voted in the last election, or the proportion of drivers who don't wear seat belts. In each of these cases, the object is to estimate a population proportion, p, using a sample proportion ($\widehat{\mathbf{P}}$).

For a Proportion, we will never know the value of the population proportion p, so we estimate p with a sample proportion. Now we will assume that we can use a normal model if $\mathbf{n\widehat{P} \geq 10}$ and $\mathbf{n(1 - \widehat{P}) \geq 10}$ as in these cases the sample is considered as a large one.

- As shown in the previous chapter, the sampling distribution of $\widehat{\mathbf{P}}$ is approximately normal.
- The mean and variance of are given by:

$$\mathbf{E}(\widehat{\mathbf{P}}) = \mathbf{P}$$

and $$\mathbf{Var}\,(\widehat{\mathbf{P}}) = \sqrt{\frac{\widehat{\mathbf{P}}(\mathbf{1}-\widehat{\mathbf{P}})}{\mathbf{n}}}$$

- The **point estimate** of $\widehat{\mathbf{P}}$ is P.
- The **100(1− α)%** confidence interval for P is

$$\widehat{\mathbf{P}} \pm \mathbf{Z}\sqrt{\frac{\widehat{\mathbf{P}}(\mathbf{1}-\widehat{\mathbf{P}})}{\mathbf{n}}}$$

Where $\widehat{\mathbf{P}}$ is the sample proportion and
n is the sample size.

Example 5.8

Suppose that a market research firm is hired to estimate the percent of adults living in a large city who have cell phones. Five hundred randomly selected adult residents in this city are surveyed to determine whether they have cell phones. Of the 500 people surveyed, 421 responded yes they own cell phones. Using a 95% confidence level, compute a confidence interval estimate for the true proportion of adult residents of this city who have cell phones.

Solution:

To calculate the confidence interval, you must find $\widehat{\mathrm{P}}$.

n = 500 , x = the number of people who said yes = 421

$\widehat{\mathrm{P}}$ = 421 ÷ 500 = 0.842

$\widehat{\mathrm{P}}$ = 0.842 is the sample proportion; this is the point estimate of the population proportion.

$\hat{\mathrm{q}}$ = 1 – $\widehat{\mathrm{P}}$ = 1 – 0.842 = 0.158

The 95% CI for the population proportion is given by:

$$95\%\ \text{CI for P} = 0.842 \pm 1.96\sqrt{\frac{0.842(0.158)}{500}} = 0.842 \pm 0.032$$

$$= [0.810\ ,\ 874]$$

Interpretation of 95% CI of the Population Proportion (P):
We estimate with 95% confidence that between 81% and 87.4% of all adult residents of this city have cell phones.

Explanation of 95% Confidence Level:
95% percent of the confidence intervals constructed in this way would contain the true value for the population proportion of all adult residents of this city who have cell phones.

Example 5.9

For a class project, a political science student at a large university wants to estimate the percent of students who are registered voters. He surveys 250 students and finds that 150 are registered voters. Compute a 99% confidence interval for the true percent of students who are registered voters, and interpret the confidence interval.

Solution:

x = 150 and n = 250

$\widehat{P}$ = x/n = 150/250 = 0.6

Since CL = 0.99, then α = 1 – CL = 1 – 0.99 = 0.01 ,

$\frac{\alpha}{2}$ = 0.005 , $Z_{\frac{\alpha}{2}}$ = 2.58

$$\text{Error (E)} = Z_{\frac{\alpha}{2}}\sqrt{\frac{\widehat{P}(1-\widehat{P})}{n}} = (2.58)\sqrt{\frac{0.6(0.4)}{250}} = 0.08$$

The confidence interval for the true population proportion is:

$$0.6 \pm 0.08 = [0.52\ ,\ 0.68].$$

Interpretation 99% CI of the Population Proportion (P):

- We estimate with 99% confidence that the true percent of all students that are registered voters is between 52% and 68%.
- Alternate Wording: We estimate with 99% confidence that between 56.4% and 63.6% of ALL students are registered voters.

Explanation of 99% Confidence Level:

99% of all confidence intervals constructed in this way contain the true value for the population percent of all students that are registered voters.

Example 5.10

A student polls his school to see if students in the school district are for or against the new school uniform. She surveys 600 students and finds that 480 are against the new uniform.

1. Compute a 95% confidence interval for the true percent of students who are against the new uniform, and interpret the confidence interval.
2. In a sample of 300 students, 68% said they own an iPod and a smart phone. Compute a 99% confidence interval for the true percent of students who own an iPod and a smartphone.

Solution:

1. To calculate the confidence interval, you must find $\hat{\mathbf{p}}$.

n = 600

x = Number of students who are against the new uniform = 480

$\hat{\mathbf{P}}$ = 480 ÷ 600 = 0.8

$\hat{\mathbf{P}}$ = 0.8 is the sample proportion; this is the point estimate of the population proportion.

The 95% CL for the population proportion is given by:

95% CL for P = $0.8 \pm 1.96\sqrt{\frac{0.8(0.2)}{600}} = 0.8 \pm 0.032$

= [0.768 , 832]

Interpretation of 95% CI of the Population Proportion (P):
We estimate with 95% confidence that the true percent of all students in the district who are against the new uniform is between 76.8% and 83.2%.

Explanation of 95% Confidence Level:
Ninety-five percent of the confidence intervals constructed in this way would contain the true value for the population proportion of all students who are against the new uniform.

2. $\widehat{P} = 0.68$

99% CI for P = $0.68 \pm 2.58\sqrt{\frac{0.68(0.32)}{300}} = 0.68 \pm 0.069$

= [0.611 , 749]

5.1.3 Determination of Sample Size:

What is ‘sample size’?

“Sample size” is a market research term used for defining the number of individuals included to conduct research. Researchers choose their sample based on demographics, such as age, gender, or physical location.

Samples can be vague or specific. For example, you may want to know what people within the 18 - 25 age range think of your product. Or, you may only require your sample to live in the Egypt, which gives you a wide range of the population. The total number of individuals in a particular sample is the sample size.

Why is it important to determine the sample size?

Are you ready to survey your research target? Research surveys help you gain insights from your target audience. The data you collect gives you insights to meet customer needs, leading to increased sales and customer loyalty. Sample size calculation and determination are imperative to the researcher to determine the right number of respondents, keeping in mind the research study's quality.

So, how should you determine the sample size? How do you know who should get your survey? How do you decide on the number of the target audience?

Sending out too many surveys can be expensive without giving you a definitive advantage over a smaller sample. But if you send out too few, you won't have enough data to draw accurate conclusions. Knowing how to calculate and determine sample size accurately can give you an edge over your competitors. Let's take a look at what a good sample includes. Also, let's look at the sample size calculation formula so you can determine the perfect sample size for your next survey.

Why do you need to determine sample size?

Let's say you are a market researcher in a city and want to send out a survey or questionnaire. The purpose of the survey is to understand your audience's feelings toward a new cell phone you are about to launch. You want to know what people in this city think about the new product to predict the phone's success or failure before launch.

Hypothetically, you choose the population of this city, which is 5 million e.g. You use a sample size determination formula to select a sample of 500 individuals that fit into the consumer panel requirement. You can use the responses to help you determine how your audience will react to the new product.

However, knowing how to determine a sample size requires more than just throwing your survey at as many people as you can. If your sample size is too big, it could waste resources, time, and money. A sample size that's too small doesn't allow you to gain maximum insights, leading to inconclusive results.

How to Calculate Sample Size?

Before we jump into sample size determination, let's take a look at the terms you should know:

1. **Confidence Level:** Confidence level tells you how sure you can be that your data is accurate. It is expressed as a percentage and aligned to the confidence interval. For example, if your confidence level is 95%, your results will most likely be 95% accurate.
2. **The Margin of Error (Confidence Interval):** When it comes to surveys, there's no way to be 100% accurate. Confidence intervals tell you how far off from the population means you're willing to allow your data to fall. A margin of error describes how close you can reasonably expect a survey result to fall relative to the real population value.
3. **Standard Deviation:** Standard deviation is the measure of the dispersion of a data set from its mean. It measures the absolute variability of a distribution. The higher the dispersion or variability, the greater the standard deviation and the greater the magnitude of the deviation. For example, you have already sent out your survey. How much variance do you expect in your responses? That variation in response is the standard of deviation.

4. Find Your Z - Score

Next, you need to turn your confidence level into a Z - score. Here are the Z-scores for the most common confidence levels:

Confidence Level	80%	85%	90%	95%	99%
z- score	1.28	1.44	1.65	1.96	2.58

If you chose a different confidence level, use our Z- score table to find your score.

5. Use the Sample Size Formula

use the sample size formula to determine the sample size.

Determination of Sample size for Estimating the Population Mean:

The module on confidence intervals provided methods for estimating confidence intervals for various parameters (e.g., μ , p, (μ_1 - μ_2), μ_d, (p_1 - p_2)). Confidence intervals for every parameter take the following general form:

Point Estimate ± Margin of Error

In the module on confidence intervals we derived the formula for the confidence interval for μ as

$$\bar{x} \pm 1.96\left(\frac{\sigma}{\sqrt{n}}\right)$$

In practice we use the sample standard deviation to estimate the population standard deviation. Note that there is an alternative formula for estimating the mean of a continuous outcome in a single population, and it is used when the sample size is small ($n < 30$). It involves a value from the t distribution, as opposed to one from the standard normal distribution, to reflect the desired level of confidence. When performing sample size computations, we use the large sample formula shown here.

Note: The resultant sample size might be small, and in the analysis stage, the appropriate confidence interval formula must be used.

The point estimate for the population mean is the sample mean and the margin of error is

$$\mathbf{Z\left(\frac{\sigma}{\sqrt{n}}\right)}$$

In planning studies, we want to determine the sample size needed to ensure that the margin of error is sufficiently small to be informative. For example, suppose we want to estimate the mean weight of female college students. We conduct a study and generate a 95% confidence interval as follows 125 $\pm$ 40 pounds, or 85 to 165 pounds. The margin of error is so wide that the confidence interval is uninformative. To be informative, an investigator might want the margin of error to be no more than 5 or 10 pounds (meaning that the 95% confidence interval would have a width (lower limit to upper limit) of 10 or 20 pounds). In order to determine the sample size needed, **the investigator must specify the desired margin of error**. It is important to note that this is not a statistical issue, but a clinical or a practical one. For example, suppose we want to estimate the mean birth weight of infants born to mothers who smoke cigarettes during pregnancy. Birth weights in infants clearly have a much more restricted range than weights of female college students. Therefore, we would probably want to generate a confidence interval for the mean birth weight that has a margin of error not exceeding 1 or 2 pounds.

The margin of error in the one sample confidence interval for μ can be written as follows:

$$\mathbf{E = Z\left(\frac{\sigma}{\sqrt{n}}\right)}$$

Our goal is to determine the sample size, n, that ensures that the margin of error, "**E**", does not exceed a specified value. We can take the formula above and, with some algebra, solve for **n**:

First, multiply both sides of the equation by the square root of n. Then cancel out the square root of n from the numerator and denominator on the right side of the equation (since any number divided by itself is equal to 1). This gives:

$$\sqrt{\mathbf{n}}\,\mathbf{E} = \mathbf{Z\sigma}$$

Now divide both sides by "E" and cancel out "E" from the numerator and denominator on the left side. This gives:

$$\sqrt{\mathbf{n}} = \frac{\mathbf{Z\sigma}}{\mathbf{n}}$$

Finally, square both sides of the equation to get:

$$\mathbf{n} = \left(\frac{\mathbf{Z\sigma}}{\mathbf{E}}\right)^{\mathbf{2}}$$

This formula generates the sample size, **n**, required to ensure that the margin of error, **E**, does not exceed a specified value. To solve for **n**, we must input "**Z**" , "**σ**" , and "**E**".

- **Z** is the value from the table of probabilities of the standard normal distribution for the desired confidence level (e.g., Z = 1.96 for 95% confidence)
- **E** is the margin of error that the investigator specifies as important from a clinical or practical standpoint.
- **σ** is the standard deviation of the outcome of interest.

Sometimes it is difficult to estimate **σ**. When we use the sample size formula above (or one of the other formulas that we will present in the sections that follow), we are **planning** a study to estimate the unknown mean of a particular outcome variable in a population. It is unlikely that we would know the standard deviation of that variable. In sample size computations, investigators often use a value for the standard deviation from a previous study or a study done in a different, but comparable, population. The sample size computation is not an application of

statistical inference and therefore it is reasonable to use an appropriate estimate for the standard deviation. The estimate can be derived from a different study that was reported in the literature; some investigators perform a small pilot study to estimate the standard deviation. A pilot study usually involves a small number of participants (e.g., n = 10) who are selected by convenience, as opposed to by random sampling. Data from the participants in the pilot study can be used to compute a sample standard deviation, which serves as a good estimate for σ in the sample size formula. Regardless of how the estimate of the variability of the outcome is derived, it should always be conservative (i.e., as large as is reasonable), so that the resultant sample size is not too small.

The formula $\mathbf{n} = \left(\frac{\mathbf{Z\sigma}}{\mathbf{E}}\right)^{\mathbf{2}}$ produces the minimum sample size to ensure that the margin of error in a confidence interval will not exceed **E**. In planning studies, investigators should also consider attrition or loss to follow - up. The formula above gives the number of participants needed with complete data to ensure that the margin of error in the confidence interval does not exceed **E**. We will illustrate how attrition is addressed in planning studies through examples in the following sections.

Example 5.11

If we are interested in estimating the amount by which a drug lowers a subject's blood pressure with a 95% confidence interval that is six units wide, and we know that the standard deviation of blood pressure in the population is 15, then how large a sample would be needed?

Solution:

The required sample size is given by

$$n = (1.96)^2 \times (15)^2 \div (3)^2 = 96.04$$

which would be rounded up to 97, because the obtained value is the minimum sample size, and sample sizes must be integers and must lie on or above the calculated minimum.

Example 5.12

A sample of 64 STAT students at a university reported they spent an average of £252.45 on textbooks and case materials per semester. Assume the population standard deviation is £74.50. How large a sample would be needed to estimate μ with a margin of error of £10 at 95% confidence?

Solution:

$Z_{\frac{\alpha}{2}} = Z_{\frac{0.05}{2}} = Z_{0.025} = 1.96$, $\sigma = 74.5$

$$n = \frac{\left(Z_{\frac{\alpha}{2}}\right)^2 \sigma^2}{(E)^2}$$

$$n = \frac{(1.645)^2(74.50)^2}{(10)^2} = 150.19 \approx 151$$

$$Z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right) = 1.645\left(\frac{74.50}{\sqrt{151}}\right) = (1.645)(6.06) \approx 10 = E$$

Example 5.13

An investigator wants to estimate the mean systolic blood pressure in children with congenital heart disease who are between the ages of 3 and 5. How many children should be enrolled in the study? The investigator plans on using a 95% confidence interval (so Z = 1.96) and wants a margin of error of 5 units. The standard deviation of systolic blood pressure is unknown, but the investigators conduct a literature search and find that the standard deviation of systolic blood pressures in children with other cardiac defects is between 15 and 20.

How large a sample should be taken to ensure the specified precision?

Solution:

To estimate the sample size, we consider the larger standard deviation in order to obtain the most conservative (largest) sample size.

$$n = \left(\frac{Z\sigma}{E}\right)^2 = \left(\frac{1.96(20)}{5}\right)^2 = 61.5$$

In order to ensure that the 95% confidence interval estimate of the mean systolic blood pressure in children between the ages of 3 and 5 with congenital heart disease is within 5 units of the true mean, a sample of size 62 (rounded to the next integer) is needed.

Note: We always round up; the sample size formulas always generate the minimum number of subjects needed to ensure the specified precision. The resulting sample size from the formula is rounded up to the next integer.

Had we assumed a standard deviation of 15, the sample size would have been n = 35. Because the estimates of the standard deviation were derived from studies of children with other cardiac defects, it would be advisable to use the larger standard deviation and plan for a study with 62 children. Selecting the smaller sample size could potentially produce a confidence interval estimate with a larger margin of error.

Example 5.14

Assuming the heights of students in a college campus are normally distributed with a standard deviation = 5 inch, find the minimum size required to construct a 99% confidence interval for mean with a maximum error = 0.5 inch.

Solution:

Given: E = 0.5 inch, σ = 5 and α = 1 – 0.95 = 0.05

Therefore, $Z_{\frac{\alpha}{2}} = Z_{0.001} = 2.58$

The formula to find the minimum sample size is

$$n = \left[\frac{(2.58 \times 5)}{0.5}\right]^2 = 665.64$$

Therefore, rounding up this value to the next integer, the minimum sample size required is 666.

Determination of Sample Size for Estimating the Population Proportion:

During an election year, we see articles in the newspaper that state **confidence intervals** in terms of proportions or percentages. For example, a poll for a particular candidate running for president might show that the candidate has 40% of the vote within three percentage points (if the sample is large enough). Often, election polls are calculated with 95% confidence, so, the pollsters would be 95% confident that the true proportion of voters who favored the candidate would be between 0.37 and 0.43: (0.40 – 0.03 , 0.40 + 0.03).

Investors in the stock market are interested in the true proportion of stocks that go up and down each week. Businesses that sell personal computers are interested in the proportion of households in a university that own personal computers. Confidence intervals can be calculated for the true proportion of stocks that go up or down each week and for the true proportion of households in this university that own personal computers.

In studies where the plan is to estimate the proportion of successes in a dichotomous outcome variable (yes/no) in a single population, the formula for determining sample size is:

$$n = \frac{P(1-P)Z^2}{E^2}$$

where **Z** is the value from the standard normal distribution reflecting the confidence level that will be used (e.g., Z = 1.96 for 95%) and **E** is the desired margin of error. p is the proportion of successes in the population. Here we are planning a study to generate a 95% confidence interval for the unknown population proportion, **p**. The equation to determine the sample size for determining p seems to require knowledge of p, but this is obviously this is a circular argument, because if we knew the proportion of successes in the population, then a study would not be necessary! What we really need is an approximate value of p or an anticipated value. The range of p is 0 to 1, and therefore the range of p(1 - p) is 0 to1. The value of p that maximizes p(1 - p) is p = 0.5. Consequently, if there is no information available to approximate p, then p = 0.5 can be used to generate the **most conservative**, or largest, sample size.

Example 5.15

An investigator wants to estimate the proportion of freshmen at his university who currently smoke cigarettes (i.e., the prevalence of smoking). How many freshmen should be involved in the study to ensure that a 95% confidence interval estimate of the proportion of freshmen who smoke is within 5% of the true proportion if:

(a) A similar study was conducted 2 years ago and found that the prevalence of smoking was 25% among freshmen.

(b) No information available about the population proportion.

Comment on your results of Parts (a) and (b).

Solution:

(a) If the investigator believes that 0.25 is a reasonable estimate of prevalence 2 years later, it can be used to plan the next

study. Using this estimate of p, the sample size needed (assuming that again a 95% confidence interval will be used and we want the same level of precision) is

$$n = \frac{P(1-P)Z^2}{E^2}$$

$$= 0.25(1 - 0.25)(1.96)^2 / (0.05)^2$$

$$= 288.12$$

Rounding up this value to the next integer, the minimum sample size required is 289.

(b) Because we have no information on the proportion of freshmen who smoke, we use 0.5 to estimate the sample size as follows:

$$n = \frac{P(1-P)Z^2}{E^2}$$

$$= 0.5(1 - 0.5)(1.96)^2 \div (0.05)^2$$

$$= 384.2$$

Rounding this value up to the next integer, the minimum sample size required is 385.

Therefore, in order to ensure that the 95% confidence interval estimate of the proportion of freshmen who smoke is within 5% of the true proportion, a sample of size 385 is needed.

Note that the answers to **Parts (a)** and **(b)** diverge, because, as expected, the conservative sample size (For p = q = 0.5) is larger than that for p = 0.25 (or any value for p other than 0.5) because the product of p and q in the numerator of the sample size formula is a maximum when p = q = 0.5.

Example 5.16

In a pilot study, n = 80 respondents have been interviewed, 56 of whom have answered "yes" to a particular survey question. What

sample size n is required if we wish to estimate the population proportion p of "yes" answers with a margin of error of 0.02 at the 99% level of confidence?

Solution:

$$\hat{P} = \frac{56}{80} = 0.7$$

$$n = \frac{P(1-P)Z^2}{E^2}$$

$$= 0.7(1 - 0.7)\ (2.58)^2 \div (0.02)^2 = 3494.61$$

The minimum sample required is 3495.

Example 5.17

Referring to the previous exercise, what sample size should we recommend if we had no pilot study data to go on? Would n be greater or less than the n recommended in the previous exercise? Assume the study is investigating a question never before researched.

Solution:

The sample size should be greater than n = 2017 since we would use p = 0.50, the most conservative value when determining sample size. In fact, when p = 0.50, n = 4161.

$$n = \frac{P(1-P)Z^2}{E^2}$$

$$= 0.5(1 - 0.5)(2.58)^2 \div (0.02)^2$$

$$= 4160.25$$

Rounding up this value to the next integer, the minimum sample size required is 4161.

5.2 Estimation Procedures for Two Populations:

5.2.1 Estimation of the Difference in Two Population Means (Large Samples):

The estimation procedures can be extended to two populations for comparative studies. For example, suppose a study is being conducted to determine differences between the salaries paid to a population of men and a population of women. Two independent simple random samples, one from the population of men and one from the population of women, would provide two sample means, $\bar{x}_1$ and $\bar{x}_2$. The difference between the two sample means, $\bar{x}_1 - \bar{x}_2$, would be used as a **point estimate** of the difference between the two population means. In other words, a **point estimate** for the difference in two population means is simply the difference in the corresponding sample means.

The sampling distribution of $\bar{x}_1 - \bar{x}_2$ would provide the basis for a confidence interval estimate of the difference between the two population means. For qualitative variables, point and interval estimates of the difference between population proportions can be constructed by considering the difference between sample proportions.

Definition: Independence

Samples from two distinct populations are independent if each one is drawn without reference to the other, and has no connection with the other.

Our goal is to use the information in the samples to estimate the difference $\mu_1 - \mu_2$ in the means of the two populations and to make statistically valid inferences about it.

Since the mean of the sample drawn from Population 1 is a good estimator of μ_1 and the mean of the sample drawn from Population 2 is a good estimator of μ_2, a reasonable **point estimate** of the difference $\mu_1 - \mu_2$ is $\mathbf{(\bar{x}_1 - \bar{x}_2)}$. In order to widen this

point estimate into a confidence interval, we first suppose that both samples are large, that is, that both $\mathbf{n_1 \geq 30}$ and $\mathbf{n_2 \geq 30}$. If so, then the following formula for a confidence interval for $\mu_1 - \mu_2$ is valid. The population standard deviations σ_1 and σ_2 are rarely known. The sample standard deviations S_1 and S_2 would be used instead of the population standard deviations. Therefore, the standard error of the difference between the two means $\mathbf{S_{(\bar{x}_1 - \bar{x}_2)}}$ is given by

$$\mathbf{S_{(\bar{x}_1 - \bar{x}_2)} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Estimation Requirements:

The approach described in this section is valid whenever the following conditions are met:

- Both samples are **simple random samples**.
- The samples are **independent**.
- Each **population** is at least 20 times larger than its respective sample.
- The sampling distribution of the difference between means is approximately normally distributed.

Generally, the sampling distribution will be approximately normally distributed when the sample size is greater than or equal to 30.

Suppose we wish to compare the means of two distinct populations. Figure 5.1 illustrates the conceptual framework of our investigation in this section. Each population has a mean and a standard deviation. We arbitrarily label one population as Population 1 and the other as Population 2, and subscript the parameters with the numbers 1 and 2 to tell them apart. We draw a random sample from Population 1 and label the sample statistics it yields with the subscript 1. Without reference to the

first sample we draw a sample from Population 2 and label its sample statistics with the subscript 2.

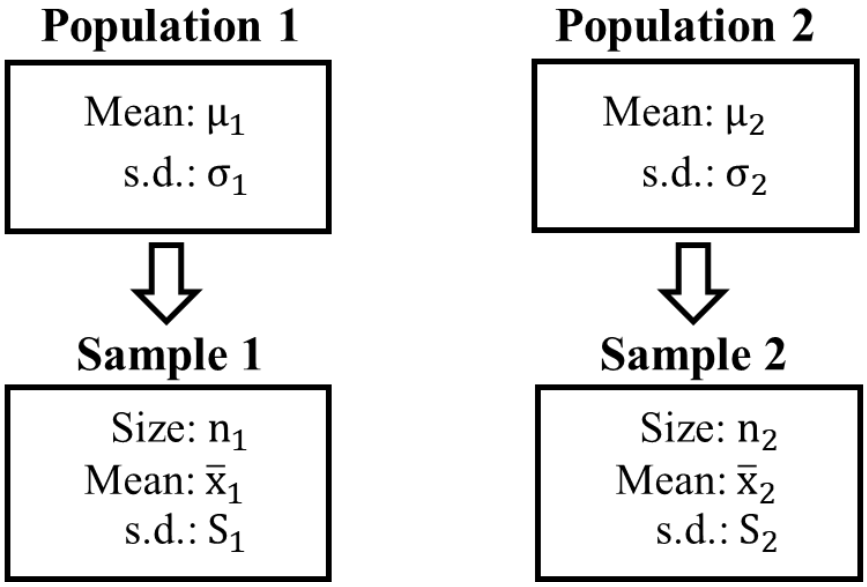


Figure 5.1: Independent Sampling from Two Populations

For large and independent samples, the 100(1− α) % confidence interval for the difference between two population means is given by

$$(\mu_1 - \mu_2) = (\bar{x}_1 - \bar{x}_2) \pm Z_{\frac{\alpha}{2}} S_{(\bar{x}_1 - \bar{x}_2)}$$

Thus, $$(\mu_1 - \mu_2) = (\bar{x}_1 - \bar{x}_2) \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

Example 5.18

To compare customer satisfaction levels of two competing cable television companies, 174 customers of Company 1 and 355 customers of Company 2 were randomly selected and were asked to rate their cable companies on a five-point scale, with 1 being least satisfied and 5 most satisfied. The survey results are summarized in the following table:

Company	Mean	Standard Deviation	Number of Customers
Company (1)	$\bar{x}_1 = 3.51$	$s_1 = 0.51$	$n_1 = 174$
Company (2)	$\bar{x}_2 = 3.24$	$s_2 = 0.52$	$n_2 = 355$

Construct a point estimate and a 99% confidence interval for $\mu_1 - \mu_2$, the difference in average satisfaction levels of customers of the two companies as measured on this five - point scale.

Solution:

The point estimate of $\mu_1 - \mu_2$ is $(\bar{x}_1 - \bar{x}_2) = 3.51 - 3.24 = 0.27$

In words, we estimate that the average customer satisfaction level for Company 1 is 0.27 points higher on this five-point scale than it is for Company 2.

The 99% confidence level means that $\alpha = 1 - 0.99 = 0.01$ so that $Z_{\frac{\alpha}{2}} = Z_{0.005} = 2.58$.

Thus,

$$(\bar{x}_1 - \bar{x}_2) \pm Z_{\frac{\alpha}{2}} S_{(\bar{x}_1 - \bar{x}_2)}$$

$$= 0.27 \pm 2.58 \sqrt{\frac{(0.51)^2}{174} + \frac{(0.52)^2}{355}}$$

$$= 0.27 \pm 0.12 = [0.15 , 0.39]$$

We are 99% confident that the difference in the population means lies in the interval [0.15 , 0.39], in the sense that in repeated sampling 99% of all intervals constructed from the sample data in this manner will contain $\mu_1 - \mu_2$. In the context of the problem we say we are 99% confident that the average level of customer satisfaction for Company 1 is between 0.15 and 0.39 points higher, on this five-point scale, than that for Company 2.

Example 5.19

Children in two elementary school classrooms were given two versions of the same test, but with the order of questions arranged from easier to more difficult in Version **A** and in reverse order in Version **B**. Randomly selected students from each class were given Version **A** and the rest Version **B**. The results are shown in the table.

Version	n	$\bar{x}$	s
A	30	85	4.6
B	36	80	4.2

Construct the 95% confidence interval for the difference in the means of the populations of all children taking Version **A** of such a test and of all children taking Version **B** of such a test.

Solution:

The point estimate of $\mu_A - \mu_B$ is $(\bar{x}_A - \bar{x}_B) = 85 - 80 = 5$

In words, we estimate that the average score for Class A is 5 points higher than it is for class B.

The 95% confidence level means that $\alpha = 1 - 0.95 = 0.05$ so that $Z_{\frac{\alpha}{2}} = Z_{0.025} = 1.96$. Thus

$$(\bar{x}_A - \bar{x}_B) \pm Z_{\frac{\alpha}{2}} S_{(\bar{x}_A - \bar{x}_B)}$$

$$= 5 \pm 1.96 \sqrt{\frac{(4.6)^2}{30} + \frac{(4.2)^2}{36}}$$

$$= 5 \pm 2.82 = [2.18 , 7.82]$$

So**,** we estimate with 95% confidence that the true difference in the means of the populations of all students taking Version **A** of such a test and of all students taking Version **B** of such a test is between 2.18 and 7.82.

We are 95% confident that the difference of population means of Version A Version B is between 2.18 and 7.82.
Based on both ends of the interval being negative, it seems like the mean of Version A is higher than the mean pf Version B.

Note: What happens if we defined $\bar{x}_A$ to be the mean of version B and $\bar{x}_B$ for the mean of version A? If you follow through the calculations, you will find that the confidence interval will differ only in sign. In other words, the interval would be - 0.782 to - 2.18. It still shows that the mean of version A is higher than the mean of version B.

Example 5.20

In comparing the academic performance of college students who are affiliated and those students who are regular, a random sample of students was drawn from each of the two populations on a university campus. Summary statistics on the student GPAs are given below.

Status	n	$\bar{x}$	s
Affiliated	40	3.0	0.4
Regular	50	3.2	0.5

Construct a point estimate and a 99% confidence interval for the difference in average GPA between the population of affiliated students and the population of regular students on this university campus.

Solution:

The point estimate of $\mu_1 - \mu_2$ is $(\bar{x}_1 - \bar{x}_2) = 3.0 - 3.2 = -0.2$
In words, we estimate that the average score for all affiliated students is 0.2 points less than it is for all regular students.

The 99% confidence level means that $\alpha = 1 - 0.99 = 0.01$ so that $Z_{\frac{\alpha}{2}} = Z_{0.025} = 2.58$. Thus

$$(\bar{x}_1 - \bar{x}_2) \pm Z_{\frac{\alpha}{2}} S_{(\bar{x}_1 - \bar{x}_2)}$$

$$= -0.2 \pm 2.58 \sqrt{\frac{(0.4)^2}{40} + \frac{(0.5)^2}{50}}$$

$$= -0.2 \pm 0.245 = [-2.045 , 0.445]$$

So, we estimate with 99% confidence that the true difference in the means of the populations of all affiliated students and of all regular students is between - 2.045 and 0.445.

Notice that you could get a negative value for $(\bar{x}_1 - \bar{x}_2)$, for example, if you had switched the affiliated and regular students, you would have gotten –1 for this difference. You would say that the mean of regular students is higher that of the affiliated students in the sample (the same conclusion stated differently).

If you want to avoid negative values for the difference in sample means, always make the group with the larger sample mean your first group.
However, even if the group with the larger sample mean serves as the first group, sometimes you will still get negative values in the confidence interval. Notice that you could get a negative value for $(\bar{x}_1 - \bar{x}_2)$.

5.2.2 Estimation of the Difference in Two Population Proportions (Large Samples):

Point Estimate:

The point estimate for the difference between the two population proportions, $P_1 - P_2$, is the difference between the two sample proportions written as $(\hat{P}_1 - \hat{P}_2)$.

We know that a point estimate is probably not a good estimator of the actual population. By adding some amount of error to this point estimate, we can create a confidence interval as we did with one sample parameters.

Interval Estimation for $P_1 - P_2$:

Consider two populations and label them as population 1 and population 2. Take a random sample of size n_1 from population 1 and take a random sample of size n_2 from population 2. If we consider them separately,

Proportion from Sample 1:

If $\mathbf{n_1\hat{P}_1 \geq 5}$ and $\mathbf{n_1(1 - \hat{P}_1) \geq 5}$,

then $\hat{\mathbf{P}}_1$ follows a normal distribution with**:**

Mean: $\mathbf{P_1}$

Standard Error: $\sqrt{\frac{P_1(1-P_1)}{n_1}}$

Estimated Standard Error: $\sqrt{\frac{\hat{P}_1(1-\hat{P}_1)}{n_1}}$

Proportion from Sample 2:

If $\mathbf{n_2\hat{P}_2 \geq 5}$ and $\mathbf{n_2(1 - \hat{P}_2) \geq 5}$,

then $\hat{\mathbf{P}}_2$ follows a normal distribution with**:**

Mean: $\mathbf{P_2}$

Standard Error: $\sqrt{\frac{P_2(1-P_2)}{n_2}}$

Estimated Standard Error: $\sqrt{\frac{\hat{P}_2(1-\hat{P}_2)}{n_2}}$

Sample Proportion 1 - Sample Proportion 2:

Using the theory introduced previously, if n_1p_1, $n_1(1 - P_1)$, n_2P_2, and $n_2(1 - P_2)$ are all greater than five and we have independent

samples, then the sampling distribution of the difference between the two sample proportions $\mathbf{(\widehat{P}_1\text{-}\widehat{P}_2)}$ is approximately normal with:

Mean: $\mathbf{P_1 - P_2}$

Standard Error: $\sqrt{\frac{\mathbf{P_1(1-P_1)}}{\mathbf{n_1}} + \frac{\mathbf{P_2(1-P_2)}}{\mathbf{n_2}}}$

Estimated Standard Error: $\sqrt{\frac{\mathbf{\widehat{P}_1(1-\widehat{P}_1)}}{\mathbf{n_1}} + \frac{\mathbf{\widehat{P}_2(1-\widehat{P}_2)}}{\mathbf{n_2}}}$

Then, we can construct the confidence interval for $P_1 - P_2$. Since we do not know P_1 and P_2, we need to check the conditions given above. If these conditions are satisfied, then the confidence interval can be constructed for two independent proportions.

The $(1-\alpha)100\%$ confidence interval of $P_1 - P_2$ is given by:

$$\mathbf{(\widehat{P}_1 - \widehat{P}_2) \pm Z\sqrt{\frac{\widehat{P}_1(1-\widehat{P}_1)}{n_1} + \frac{\widehat{P}_2(1-\widehat{P}_2)}{n_2}}}$$

which, in interval notation, is:

$$\mathbf{(\widehat{P}_1 - \widehat{P}_2) - Z_{\frac{\alpha}{2}}\sqrt{\frac{\widehat{P}_1(1-\widehat{P}_1)}{n_1} + \frac{\widehat{P}_2(1-\widehat{P}_2)}{n_2}}, (\widehat{P}_1 - \widehat{P}_2) + Z_{\frac{\alpha}{2}}\sqrt{\frac{\widehat{P}_1(1-\widehat{P}_1)}{n_1} + \frac{\widehat{P}_2(1-\widehat{P}_2)}{n_2}}}$$

Where $\mathbf{\widehat{P}_1}$ and n_1 are the sample proportion and sample size of the first sample, and $\mathbf{\widehat{P}_2}$ and n_2 are the sample proportion and sample size of the second sample. The value $\mathbf{Z_{\frac{\alpha}{2}}}$ is the appropriate value from the standard normal distribution for your desired confidence level. (Refer to the table for $\mathbf{Z_{\frac{\alpha}{2}}}$-values given before).

The formula shown here for a confidence interval for $(P_1 - P_2)$ is used under the condition that both of the sample sizes are large enough for the Central Limit Theorem to be applied and allow you to use a $\mathbf{Z_{\frac{\alpha}{2}}}$ value; this is true when you are estimating proportions

using large scale surveys, for example. For small sample sizes, confidence intervals are beyond the scope of this course.

Statistical Significance and Confidence Intervals:

- **If the two confidence intervals do not overlap**, we can conclude that there is a statistically significant difference in the two population values at the given level of confidence; or alternatively
- **If the confidence interval for the difference does not contain zero**, we can conclude that there is a statistically significant difference in the two population values at the given level of confidence.

The first rule is the "more conservative" one since there are some circumstances when the interval for the difference does not contain zero but there is some overlap in the individual confidence intervals.

Importantly, the formula for the standard error of a difference is for two independent samples. It would not apply to dependent samples like those gathered in a matched pairs study.

Example 5.21

Males and females were asked about what they would do if they received a $100 bill by mail, addressed to their neighbor, but wrongly delivered to them. Would they return it to their neighbor? Of the 100 males sampled, 85 said "yes" and of the 120 females sampled, 96 said "yes."

Find a 95% confidence interval for the difference in proportions for males and females who said "yes."

Solution:

Let sample one be males and sample two be females. Then we have:

Males:

$n_1 = 100$, $\widehat{P}_1 = 85/100 = 0.85$

Females:

$n_2 = 120$, $\widehat{P}_2 = 96/120 = 0.80$

Checking conditions we see that $n_1\widehat{P}_1$, $n_1(1 - \widehat{P}_1)$, $n_2\widehat{P}_2$, and $n_2(1-\widehat{P}_2)$ are all greater than five so our conditions are satisfied.

Using the formula above, we get:

$$(\widehat{P}_1 - \widehat{P}_2) \pm Z\sqrt{\frac{\widehat{P}_1(1-\widehat{P}_1)}{n_1} + \frac{\widehat{P}_2(1-\widehat{P}_2)}{n_2}}$$

$$= (0.85 - 0.80) \pm 1.96\sqrt{\frac{0.85(0.15)}{100} + \frac{0.8(0.2)}{120}}$$

$$= 0.05 \pm 0.10 = [-0.05 , 0.15]$$

We are 95% confident that the difference of population proportions of males who said "yes" and females who said "yes" is between -0.05 and 0.15.

Based on both ends of the interval being negative and positive, this interval contains the value zero, so it seems like the difference between the proportion of females who would return it and the proportion of males who would return it is not statistically significant.

Note: What happens if you had switched the males and females? If you follow through the calculations, you will find that the confidence interval will differ only in sign. In other words, if female was $\widehat{P}_1$, the interval would be -0.15 to 0.05. It still shows that the proportion of females and the proportion of males are equal.

Example 5.22

A medical researcher anticipates that smoking can result in the heart diseases. The researcher recruited **150 smokers** and **250**

nonsmokers to take part in an observational study and found that **95** of the **smokers** and **105** of the **nonsmokers** were seen to have heart diseases.
Construct a point estimate and a **99%** confidence interval of the difference between the two population proportions.

Solution:

Smokers:

$n_1 = 150$, $x_1 = 95$, $\widehat{P}_1 = 95 \div 150 = 0.63$,

Standard Error = $\sqrt{\frac{0.63(0.37)}{150}} = 0.0394$

Nonsmokers:

$n_2 = 250$, $x_2 = 105$, $\widehat{P}_2 = 105 \div 250 = 0.42$,

Standard Error = $\sqrt{\frac{0.42(0.58)}{250}} = 0.0312$

Smokers – Nonsmokers

Point Estimate:

The difference between the two sample proportions is:
0.63 – 0.42 = 0.21 which is a point estimate of the difference between the two population proportions.

Interval Estimation:

We would like to make a confidence interval for the true difference that would exist between these two groups in the population.
So we compute:

Standard Error for Difference = $\sqrt{(0.0394)^2 + (0.0312)^2} = 0.05$

If we think about all possible ways to draw a sample of 150 smokers and 250 non-smokers then the differences we'd see between sample proportions would approximately follow the normal distribution. Thus, a 95% Confidence Interval for the

differences between these two proportions in the population is given by:

$$0.21 \pm 2.58(0.05) \text{ or } 0.21 \pm 0.13 = [0.08 , 0.34]$$

Interpretation: To interpret these results within the context of the problem, we are 95% confident that the difference in the proportion of heart diseases in smokers as compared to non-smokers is between 0.08 and 0.34. The **null value** for the risk difference is zero. Because the 95% confidence interval does not include zero, we conclude that a higher percentage of smokers than nonsmokers have seen to have heart diseases and the difference in prevalent heart diseases between smokers and non-smokers is statistically significant.

If you had switched the smokers and nonsmokers, you would have gotten -0.21 for this difference. That's okay, but you can avoid negative differences in the sample proportions by having the group with the larger sample proportion serve as the first group (here, smokers).

Another way to think about whether the smokers and non-smokers have significantly different proportions with heart diseases is to calculate a 95% confidence Interval for each group separately. For the smokers, we have a confidence interval of $0.63 \pm 2(0.0394)$ or 0.63 ± 0.0788. The interval for smokers goes from about 0.55 up to 0.71. For the non-smokers, we have a confidence interval of $0.42 \pm 2(0.0312)$ or 0.42 ± 0.0624. The interval for non-smokers goes from about 0.36 up to 0.48. The interval for the smokers (which starts at 0.55) and the interval for the non-smokers (which ends at 0.48) do not overlap. that is, the two population proportions are significantly different.

Example 5.23

How much more likely are women than men over 80 to develop Alzheimer's disease? 90 of the 900 men over 65 years old

observed developed Alzheimer's disease and 300 of the 1200 women over 80 years old observed developed Alzheimer's disease. Come up with and interpret the 99% confidence interval for the difference.

Solution:

Men:

$n_1 = 900$, $x_1 = 225$, $\widehat{P}_1 = 225 \div 900 = 0.25$,

Standard Error = $\sqrt{\frac{0.25(0.75)}{900}} = 0.0144$

Women:

$n_2 = 1200$, $x_2 = 120$, $\widehat{P}_2 = 120 \div 1200 = 0.1$,

Standard Error = $\sqrt{\frac{0.1(0.9)}{1200}} = 0.0087$

Men – Women

Point Estimate:

The difference between the two sample proportions is:

0.25 – 0.10 = 0.15 which is a point estimate of the difference between the two population proportions.

Interval Estimation:

We would like to make a confidence interval for the true difference that would exist between these two groups in the population.

So we compute:

Standard Error for Difference = $\sqrt{(0.0144)^2 + (0.0087)^2} = 0.0168$

If we think about all possible ways to draw a sample of 900 men and 1200 women then the differences we would see between sample proportions would approximately follow the normal distribution. Thus, a 95% Confidence Interval for the differences between these two proportions in the population is given by:

$0.15 \pm 2.58(0.0168)$ or $0.15 \pm 0.04 = [0.11 , 0.19]$

Interpretation: To interpret these results within the context of the problem, we are 99% confident that the difference in the proportion of Alzheimer's in men as compared to women is between 0.11 and 0.19. The **null value** for the risk difference is zero. Because the 99% confidence interval does not include zero, we conclude that a higher percentage of men than women have seen to have Alzheimer's and the difference in prevalent Alzheimer's between men and women is statistically significant.

The 99% confidence interval is [0.11 , 0.19]. With 95% confidence it can be concluded that for the population of all men and women over 80 years old, that men are between 11% and 19% greater likely than women to develop Alzheimer's.

Note: If you had switched men and women, the 99% confidence interval for the difference between women and men becomes [- 0.19 , - 0.11]. If this is the required case, you can avoid negative differences in the sample proportions by having the group with the larger sample proportion serve as the first group (here, men).

Hint: Since the two topics “Estimation” and “Hypothesis Testing” are intertwined and built on a common ground, we will present the “Exam Questions” together at the end of Chapter 6.

Chapter (6)
Hypothesis Testing

Contents

Chapter 6
Hypothesis Testing

6.1 Introduction:

One job of a statistician is to make statistical inferences about populations based on samples taken from the population. Confidence intervals are one way to estimate a population parameter. Another way to make a statistical inference is to make a decision about a parameter. For instance, a car dealer advertises that its new small truck gets 35 miles per gallon, on the average. A tutoring service claims that its method of tutoring helps 90% of its students get an A or a B.

A company says that women managers in their company earn an average of $60,000 per year.

A statistician will make a decision about these claims. This process is called "hypothesis testing." A hypothesis test involves collecting data from a sample and evaluating the data. Then, the statistician makes a decision as to whether or not the data supports the claim that is made about the population.

In this chapter, you will conduct hypothesis tests for one population and two populations will be considered.

Definition: The Hypothesis is an assumption which is tested to check whether the inference drawn from the sample of data stand true for the entire population or not.
Hypothesis testing is a form of statistical inference that uses data from a sample to draw conclusions about a population parameter. First, a tentative assumption is made about the parameter. This assumption is called the **null hypothesis** and is denoted by $\mathbf{H_0}$. An **alternative hypothesis** (denoted by $\mathbf{H_a}$), which is the opposite of what is stated in the null hypothesis, is then defined. The null and alternative hypotheses

are mutually exclusive, and only one can be true. However, one of the two hypotheses will always be true.

The hypothesis-testing procedure involves using sample data to determine whether or not H_0 can be rejected. If H_0 is rejected, the statistical conclusion is that the alternative hypothesis H_a is true.

Null Hypothesis: The hypothesis to be tested, denoted H_o. It is a statement about the population that will be assumed to be true until it is declared false beyond a reasonable doubt. (Null hypothesis contains the equal sign).

Alternative Hypothesis: It is a claim about the population that is contradictory to H_0 and what we conclude when we reject H_0 (denoted by H_a). A hypothesis considered to be an alternate to the null hypothesis, denoted H_a. (Alternative hypothesis contains an inequality $\neq$, < , and >).

You might notice that **we don't say that we reject or fail to reject the alternate hypothesis**. This is because hypothesis testing is not designed to prove or disprove anything. It is only designed to test whether a pattern we measure could have arisen spuriously, or by chance.

After you have determined which hypothesis the sample supports, you make a decision. There are two options for a decision. They are "reject H_0" if the sample information favors the alternate hypothesis or "do not reject H_0" if the sample information favors the null hypothesis, meaning that there is not enough information to reject the null hypothesis.

Mathematical Symbols Used in H_0 and Ha

H_0	H_a
equal (=)	not equal ($\neq$) or greater than (>) or less than (<)
greater than or equal to ($\geq$)	less than (<)
Less than or equal to ($\leq$)	Greater than (>)

NOTE: H_0 always has a symbol with an equal in it. H_a never has a symbol with an equal in it. The choice of symbol depends on the wording of the hypothesis test. However, be aware that many statisticians use (=) in the Null Hypothesis, even with > or < as the symbol in the Alternate Hypothesis. This practice is acceptable because we only make the decision to reject or not reject the Null Hypothesis.

Real Examples of Hypothesis Testing:

- If, for example, a person wants to test that a coin has exactly a 50% chance of landing on heads, the null hypothesis would be that 50% is correct, and the alternative hypothesis would be that 50% is not correct.

Statistically, the null hypothesis would be represented as H_0: P = 0.5. The alternative hypothesis would be denoted as "H_a" and be identical to the null hypothesis, except with the equal sign, meaning that it does not equal 50%.

A random sample of 100 coin flips is taken, and the null hypothesis is then tested. If it is found that the 100 coin flips were distributed as 40 heads and 60 tails, the analyst would assume that a coin does not have a 50% chance of landing on heads and would reject the null hypothesis and accept the alternative hypothesis.

If, on the other hand, there were 48 heads and 52 tails, then it is reasonable that the coin could be fair and still produce such a result. In cases such as this where the null hypothesis is "accepted," the analyst states that the difference between the expected results (50 heads and 50 tails) and the observed results (48 heads and 52 tails) is "explainable by chance alone".

- Here is a simple example. A school principal claims that students in her school score an average of 7 out of 10 in exams. The null hypothesis is that the population mean is 7.0. To test this

null hypothesis, we record marks of say 30 students (sample) from the entire student population of the school (say 300) and calculate the mean of that sample.
We can then compare the (calculated) sample mean to the (hypothesized) population mean of 7.0 and attempt to reject the null hypothesis.

For the above example, **null hypotheses** is:

Students in the school score an average of 7 out of 10

For the purposes of determining whether to reject the null hypothesis, the null hypothesis (abbreviated H_0) is assumed, for the sake of argument, to be true. Then the likely range of possible values of the calculated statistic (e.g., the average score on 30 students' tests) is determined under this presumption (e.g., the range of plausible averages might range from 6.2 to 7.8 if the population mean is 7.0). Then, if the sample average is outside of this range, the null hypothesis is rejected. Otherwise, the difference is said to be "explainable by chance alone," being within the range that is determined by chance alone.
An important point to note is that we are testing the null hypothesis because there is an element of doubt about its validity. Whatever information that is against the stated null hypothesis is captured in the alternative hypothesis **(H_1)**.
For the above examples, the **alternative hypothesis** would be:

Students score an average that is not equal to 7

In other words, the **alternative hypothesis** is a direct contradiction of the null hypothesis.

Ideally, the hypothesis - testing procedure leads to the acceptance of H_0 where H_0 is true and the rejection of H_0 when H_0 is false. Unfortunately, since hypothesis tests are based on sample information, the possibility of errors must be considered. A type I error corresponds to rejecting H_0 when H_0 is actually true, and

a type II error corresponds to accepting H_0 when H_0 is false. The probability of making a type I error is denoted by α, and the probability of making a type II error is denoted by β.

In using the hypothesis testing procedure to determine if the null hypothesis should be rejected, the person conducting the hypothesis test specifies the maximum allowable probability of making a type I error, called the level of significance for the test. Common choices for the level of significance are $\alpha = 0.05$ and $\alpha = 0.01$. Although most applications of hypothesis testing control the probability of making a type I error, they do not always control the probability of making a type II error.

Hypothesis Testing Procedure:

1. **State the Null and Alternative Hypotheses:** The first step is to establish the hypothesis to be tested. The statistical hypothesis is an assumption about the value of some unknown parameter, and the hypothesis provides some numerical value or range of values for the parameter. Here two hypotheses about the population are constructed **Null Hypothesis** and **Alternative Hypothesis**.

   The Null Hypothesis denoted by H_0 asserts that there is no true difference between the sample of data and the population parameter and that the difference is accidental which is caused due to the fluctuations in sampling. Thus, a null hypothesis states that there is no difference between the assumed and actual value of the parameter.

   The alternative hypothesis denoted by H_1 is the other hypothesis about the population, which stands true if the null hypothesis is rejected. Thus, if we reject H_0 then the alternative hypothesis H_1 gets accepted.

2. **Select the Level of Significance:**
   The significance level is used as a basis to determine the rejection region, since it is the probability of rejecting a true null hypothesis or in other words the probability the test statistic will fall in the rejection region when in fact the null hypothesis is true.
   Once the hypothesis about the population is constructed the researcher has to decide the level of significance, i.e. a confidence level with which the null hypothesis is accepted or rejected. The significance level is denoted by **'α'** and is usually defined before the samples are drawn such that results obtained do not influence the choice. In practice, we either take 5% or 1% level of significance.

   If the 5% level of significance is taken, it means that there are five chances out of 100 that we will reject the null hypothesis when it should have been accepted, i.e. we are about 95% confident that we have made the right decision. Similarly, if the 1% level of significance is taken, it means that there is only one chance out of 100 that we reject the hypothesis when it should have been accepted, and we are about 99% confident that the decision made is correct.

3. **Calculate the Test Statistic:** Test statistic is the statistic used as a basis for deciding whether the null hypothesis should be rejected. If the test statistic results in a value that is in the rejection region we will reject the null hypothesis, H_0. If the test statistic results in a value that is not in the rejection region we will accept the null hypothesis. After the hypothesis are constructed, and the significance level is decided upon, the next step is to determine a suitable test statistic and its distribution. Most of the statistic tests assume the following form:

$$\text{Test Statistic} = \frac{\text{Statistic} - \text{Parameter}}{\text{Standard Deviation of Statistic}}$$

4. **Determining the Rejection Region (R.R.):** Before the samples are drawn it must be decided that which **values to the test statistic** will lead to the acceptance of H_0 and which will lead to its rejection. The values that lead to rejection of H_0 is called the critical region.
   Rejection Region (R.R.): The set of values for the test statistic that leads to the rejection of H_0.
5. **Make a Decision:** We apply the formula of the test statistic as shown in **step (3)** to check whether the sample results fall in the rejection or non-rejection regions. The statistical conclusions can be drawn, and the management can take decisions. The decision involves either accepting the null hypothesis or rejecting it. The decision that the null hypothesis is accepted or rejected depends on whether the computed value falls in the acceptance region or the rejection region.
6. **Conclusion**: Interpret the result of the hypothesis test. That is, conclusion is written in terms of the original problem.

Thus, to test the hypothesis, it is necessary to follow these steps systematically so that the results obtained are accurate and do not suffer from either of the statistical error Viz. Type - I error and Type - II error.

6.2 Testing for the Population Mean: Large Samples

If the population standard deviation (σ) is known, it is used for calculating the value of the test - statistic (S). If it is unknowing, the sample standard deviation is used.

Step 1: State the null and alternative hypotheses.

Null Hypothesis. $\mathbf{H_0 : \mu = \mu_0}$ (where μ_0 is a specified value).

Alternative Hypothesis.

- **$H_a : \mu \neq \mu_0$ (two - tailed test).**
- **$H_a : \mu > \mu_0$ or $\mu < \mu_0$ (one - tailed test).**

Step 2: State the level of significance (α): Usually 0.05 or 0.01.

Step 3: Calculate the test statistic: $Z = \dfrac{\bar{x} - \mu}{\dfrac{\sigma}{\sqrt{n}}}$

Step 4: Determine Rejection Region:

one - tailed (>): Reject H_0 if $Z > Z_\alpha$

two - tailed (≠): Reject H_0 if $Z > Z_{\frac{\alpha}{2}}$ or $Z < -Z_{\frac{\alpha}{2}}$

Table 6.1
Critical z-values

Test	$\alpha = 0.05$		$\alpha = 0.01$		$\alpha = 0.10$	
One - tailed	(>) +1.645	(<) - 1.645	(>) +2.33	(<) - 2.33	(>) +1.28	(<) - 1.28
Two - tailed	± 1.96		± 2.58		± 1.645	

Step 5: Make a Decision by determining if the calculated test statistic is in the critical rejection region or not (Reject or do not reject H_0).

The following Figures show the rejection and nonrejection regions for one - tailed and two - tailed tests.

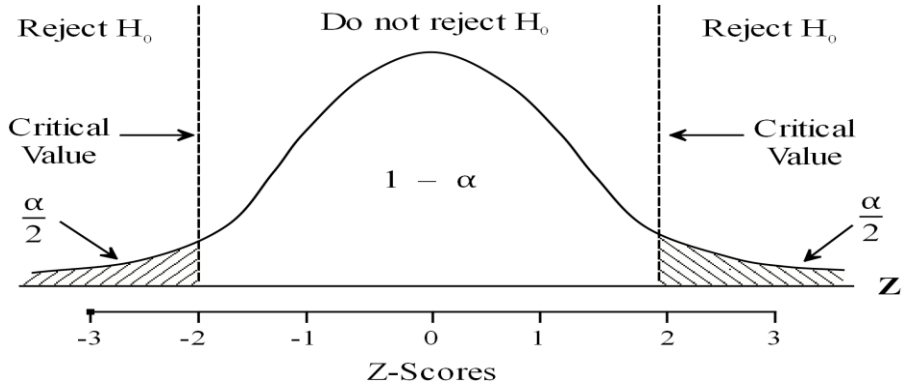


Figure 6.1
Rejection Regions for a Two-Tailed Test

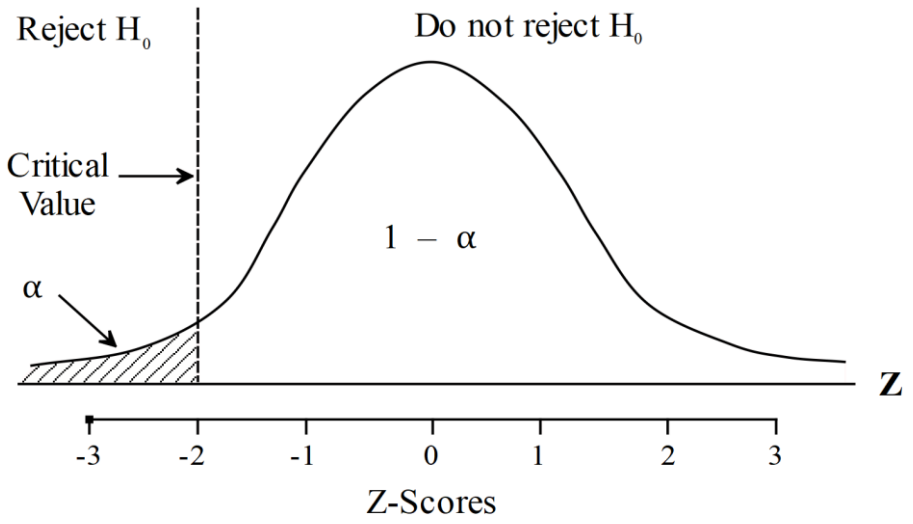


Figure 6.2
Rejection Region for a Left-Tailed Test

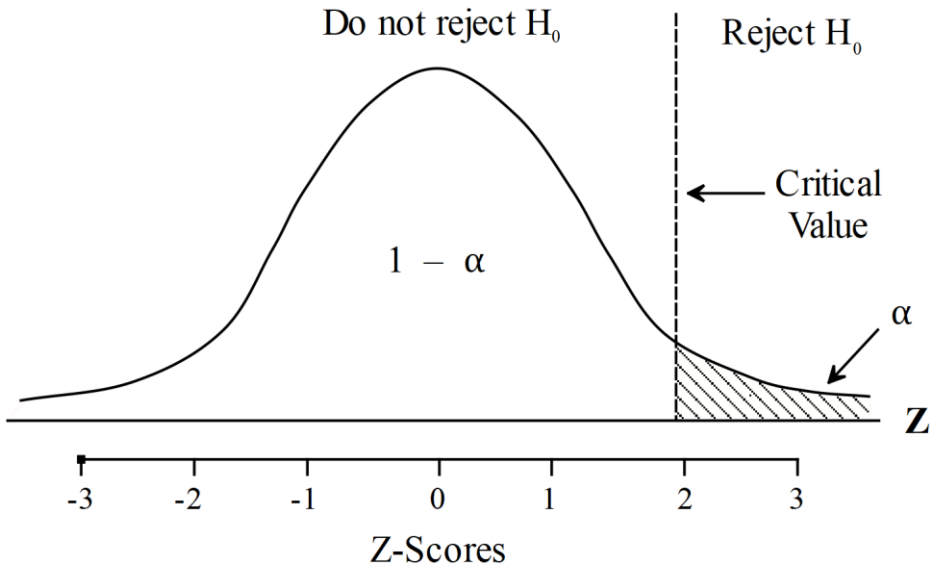


Figure 6.3
Rejection Regions for a Right-Tailed Test

The following three figures (6.3, 6.4 and 6.5) show the critical values for rejecting the null hypothesis for one-tailed and two-tailed tests and $\alpha = 0.05$:

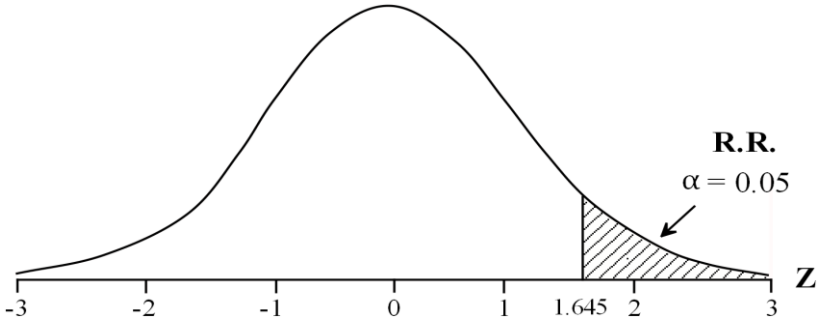


Figure 6.4
The critical Value for a Right -Tailed test and α = 0.05

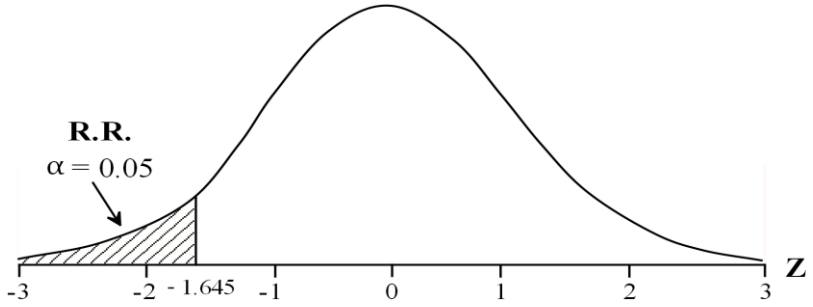


Figure 6.5
The critical Value for a Left -Tailed test for α = 0.05

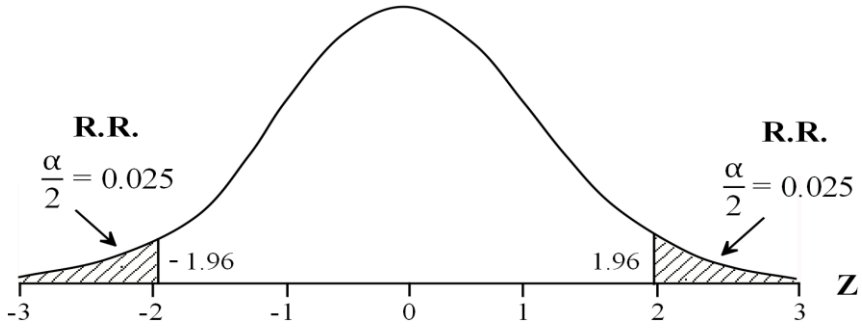


Figure 6.6
The critical Value for a Two-Tailed test and α = 0.05

Example 6.1

Suppose that babies in the town had a mean birth weight of 3,500 grams in 2010. This year, a random sample of 50 babies has a mean weight of about 3,400 grams with a standard deviation of about 500 grams. Here is the distribution of birth weights in the sample.

Does this sample give strong evidence that the town's mean birth weight is less than 3,500 grams this year? Use a significance level of 5%.

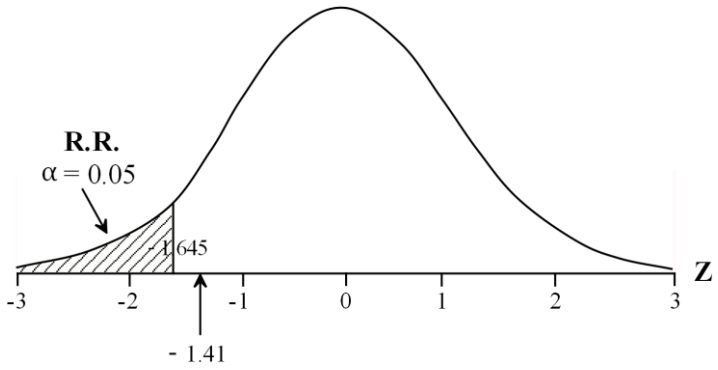
Let μ = mean birth weight in the town this year.
The null hypothesis says there is "no change from 2010".

- **H_0: μ = 3,500**
- **H_a: μ < 3,500**

Since the sample is large, we can conduct the Z - test (without worrying about the shape of the distribution of birth weights for individual babies).

$$Z = \frac{(3{,}400 - 3{,}500)}{\frac{500}{\sqrt{50}}} = -1.41$$

For the left - tailed test and significance level α = 0.05, the value of z_α is -1.645.
Since the value of the test statistic z is in the non - rejection region, we fail to reject the null hypothesis.



Conclusion: This sample does not suggest that the mean birth weight this year is less than 3,500 grams. The sample from this year has a mean of 3,400 grams, which is 100 grams lower than the mean in 2010. But this difference is not statistically significant. It can be explained by the chance fluctuation we expect to see in random sampling.

Example 6.2

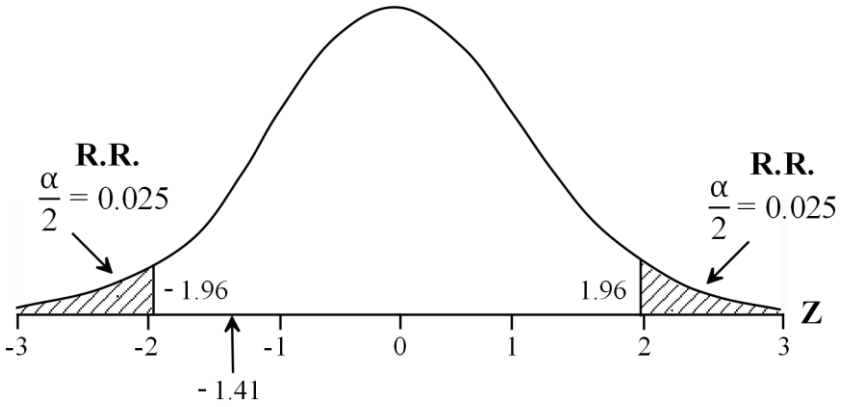
For sample data given in **Example 6.1**, Does this sample give strong evidence that the town's mean birth weight is different from 3,500 grams this year?

The null and alternative hypotheses in this case will be:

- **H_0: $\mu = 3{,}500$**
- **H_a: $\mu \neq 3{,}500$**

The value of the test statistic was found to be -1.41.

For the significance level $\alpha = 0.05$, the value of $\mathbf{Z}_{\frac{\alpha}{2}}$ is ± 1.96 as the test is a two-tailed-test.



Since the value of the test statistic z is in the non-rejection region, we still fail to reject the null hypothesis.

conclusion: No change in our conclusion.

Example 6.3

A researcher believes that there has been a reduction in the mean number of hours that college students spend preparing for final exams. A national study stated that students at a 4-year college spend an average of 25 hours preparing for 5 final exams each semester with a population standard deviation of 8 hours. The researcher sampled 100 students and found a sample mean study time of 27 hours. Does this indicate that the average study time

for final exams has increased? Use a 1% level of significance to test this claim.

Solution:

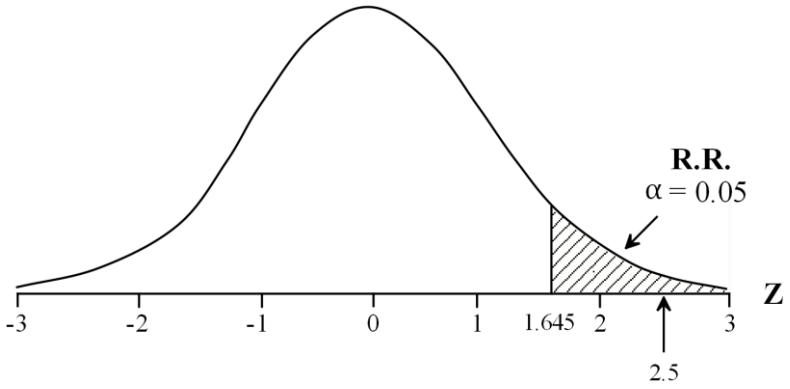
State the null and alternative hypotheses.

- **H_0: μ = 25** hours
- **H_1: μ > 25** hours

The value of the test statistic is:

$$Z = \frac{(27 - 25)}{\frac{8}{\sqrt{100}}} = 2.5$$

For the significance level α = 0.01, the value of $\mathbf{z_\alpha}$ is 2.33 as it is a right-tailed test.



The critical value is 2.33. This value sets up the rejection region. Comparing the value of test statistic to the critical value shows that the test statistic falls in the rejection region. The test statistic of 2.5 is greater than the critical value of 2.33.

We reject the null hypothesis. We have sufficient evidence to support the claim that the mean final exam study time has increased above 25 hours.

Conclusion: This sample suggests that the average study time for final exams has increased. The sample has a mean of 27 hours, which is 2 hours greater than the mean of students at this collage, and this difference is statistically significant.

6.3 Testing for a Population Proportion: (Large Samples)

The single proportion (or one-sample) is used to compare a proportion of responses or values in a sample of data to a (hypothesized) proportion in the population from which our sample data are drawn. This is important because we seldom have access to data for an entire population.

We can perform either a one-sided test (i.e., less than or greater than) or a two-tailed test (i.e., not equal). We use one-sided tests to evaluate if the available data provide evidence that a sample proportion is larger (or smaller) than the comparison value (i.e., the population value in the null - hypothesis).

To test a population proportion, there are a few things that need to be defined first. Usually, Greek letters are used for parameters and Latin letters for statistics. When talking about proportions, it makes sense to use p for proportion. The Greek letter for p is π, but that is too confusing to use. Instead, it is best to use p for the population proportion. That means that a different symbol is needed for the sample proportion. The convention is to use, $\widehat{\mathbf{P}}$, known as p-hat. This way you know that p is the population proportion, and that $\widehat{\mathbf{P}}$ is the sample proportion related to it.

Now proportion tests are about looking for the percentage of individuals who have a particular attribute. You are really looking for the number of successes that happen.

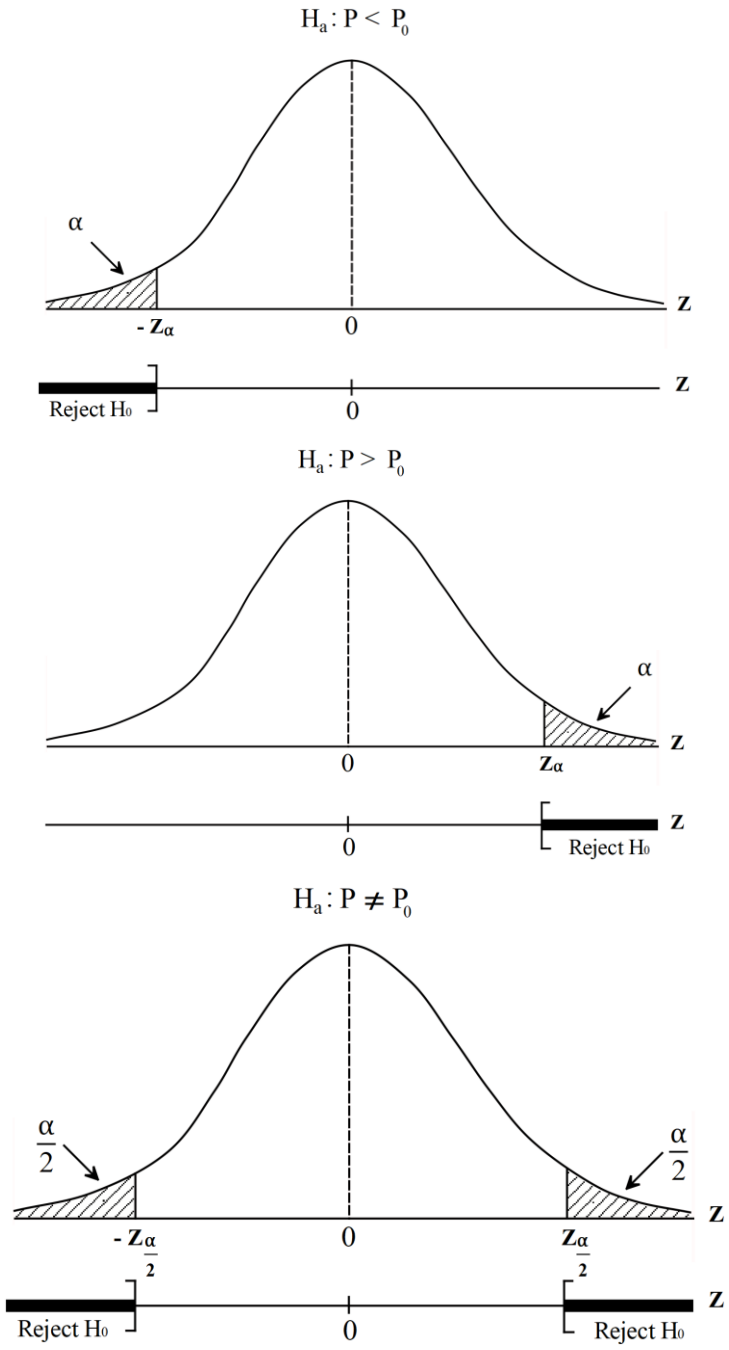
1. State the null and alternative hypotheses and the level of significance:

   $\mathbf{H_0: P = P_o}$, where po is the known proportion.

   $\mathbf{Ha: P < P_o}$, $\mathbf{H_a: P > P_o}$, or $\mathbf{H_a: P \neq P_o}$

   Use the appropriate one for your problem.

   Also, state your α level here (0.05, 0.01 or 0.10).



2. To determine the sampling distribution of $\hat{\mathbf{P}}$, you need to show that $\mathbf{n\hat{P} \geq 5}$ and $\mathbf{n\hat{q} \geq 5}$, where $\mathbf{q = 1 - p}$. If this requirement is true, then the sampling distribution of $\hat{\mathbf{P}}$ is well approximated by a normal curve.

3. Find the sample statistic and test statistic.

Sample Proportion:

$$\widehat{P} = \frac{x}{n}$$

Test Statistic: $$Z = \frac{\left(\widehat{P} - P_0\right)}{\sqrt{\frac{P_0 q_0}{n}}}$$

4. Determine Rejection Region:

one - tailed (>): Reject H_0 if $Z > Z_{\frac{\alpha}{2}}$

two - tailed (≠): Reject H_0 if $Z > Z_{\frac{\alpha}{2}}$ **or** $Z < -Z_{\frac{\alpha}{2}}$

5. Make a Decision by determining if the calculated test statistic is in the critical rejection region or not (Reject or do not reject H_0).

6. Conclusion: This is where you interpret in real world terms the conclusion to the test. The conclusion for a hypothesis test is that you either have enough evidence to show $\mathbf{H_a}$ is true, or you do not have enough evidence to show $\mathbf{H_a}$ is true.

Example 6.4

In 2015, 40% of adults aged 18 years or older reported that they had “a great deal” of confidence in the public schools. On June 1, 2020, an organization released results of a poll in which 372 of 1004 adults aged 18 years or older stated that they had “a great deal” of confidence in public schools. Does the evidence suggest at the α = 0.05 significance level that the proportion of adults aged 18 years or older having “a great deal” of confidence in the public schools is significantly lower in 2015 than the 2015 proportion?

Solution:

Step 1: State the null and alternative hypotheses.

Basically, the goal of this problem is to see whether attitudes about public schooling have changed over time. We are asked to use the results from 2015 as the "baseline" and see whether, ten years later, attitudes are lower. Thus:

- **H_0: p = 0.40**
- **H_a: p < 0.40**

Notice that this is a one-tail test since the question in the example wants to know whether confidence levels are lower.

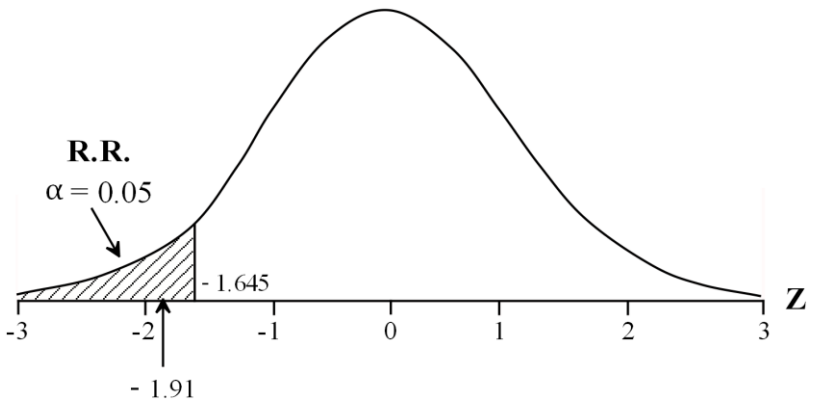
Step 2: Determine the level of significance.
We are asked to use $\alpha = 0.05$.

Step 3: Calculate the test statistic.
We first need to identify the sample proportion and standard deviation from the information given in the problem. We see that:

$$\widehat{P} = \frac{x}{n} = \frac{\mathbf{372}}{\mathbf{1004}} \approx 0.3705$$

Using this information, the value of the test statistic is:

$$Z = \frac{\widehat{P} - P_0}{\sqrt{\frac{P_0(1 - P_0)}{n}}} = \frac{0.3705 - 0.40}{\sqrt{\frac{0.4(1 - 0.4)}{1004}}} \approx -\frac{0.0295}{0.015461} \approx -1.91$$



Step 4: **Make a decision:** The value of the test statistic falls in the rejection region. Therefore, the decision is to reject the null hypothesis and accept the alternative hypothesis.

Step 5: **Make appropriate conclusions:**

Thus, **we reject the null hypothesis**, H_0: p = 0.40. Our sample data provide significant evidence that the population proportion is not 0.40, and in fact, is likely much less. This means that significantly fewer people had “a great deal” of confidence in public schools in the year 2020 compared with the year 2015.

Example 6.5

Newborn babies are more likely to be boys than girls. A random sample found 13,173 boys were born among 25,468 newborn children. Is this sample evidence that the birth of boys is more common than the birth of girls in the entire population?

Solution:

Here, we want to test: **H_0: p = 0.5** , **H_a: p > 0.5**

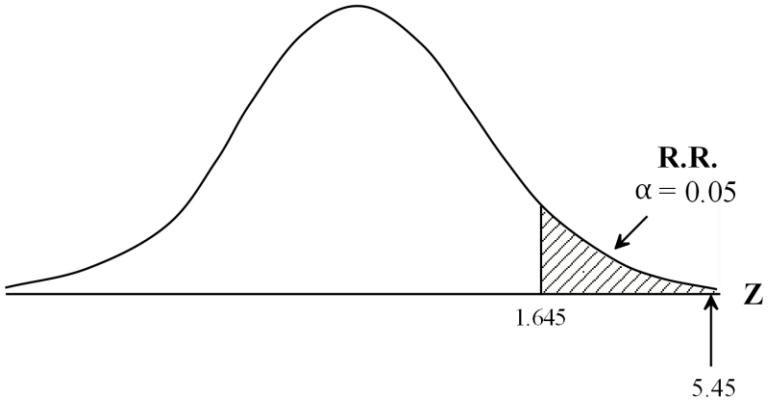
The test statistic:

$$\widehat{P} = \frac{x}{n} = \frac{\mathbf{13{,}173}}{\mathbf{25{,}468}} = 0.517$$

Test Statistic:

$$Z = \frac{\left(\widehat{P} - P_0\right)}{\sqrt{\frac{P_0 q_0}{n}}} = \frac{(0.517 - 0.5)}{\sqrt{\frac{0.5(0.5)}{25{,}468}}} = 5.43$$

Here's a picture of such a "critical region" (or "rejection region"):



Since the value of the test statistic falls in the rejection region, we will reject the null hypothesis and accept the alternative hypothesis. or equivalently since our test statistic Z = 5.49 is greater than 1.645.

Our Conclusion: We say there is sufficient evidence to conclude boys are more common than girls in the entire population.

Example 6.6

A soft drink maker claims that a majority of adults prefer its leading beverage over that of its main competitor's. To test this claim 500 randomly selected people were given the two beverages in random order to taste. Among them, 270 preferred the soft drink maker's brand, 211 preferred the competitor's brand, and 1919 could not make up their minds. Determine whether there is sufficient evidence, at the 5% level of significance, to support the soft drink maker's claim against the default that the population is evenly split in its preference.

Solution:

The null and alternative hypotheses are:

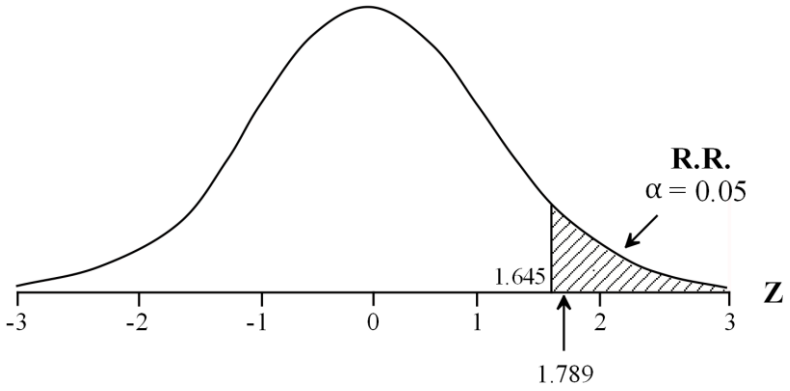
$$\mathbf{H_0: p = 0.50 \quad , \quad H_a: p > 0.50}$$

where p denotes the proportion of all adults who prefer the company's beverage over that of its competitor's beverage.

The test statistic

$$Z = \frac{(\hat{P} - P_0)}{\sqrt{\frac{P_0 q_0}{n}}} = \frac{(0.54 - 0.5)}{\sqrt{\frac{0.5(0.5)}{500}}} = 1.7981$$

For α = 0.05, the value of $\mathbf{Z_\alpha}$ is 1.645 as the test is a right - tailed test.



As shown in this Figure, the test statistic falls in the rejection region.

The decision is to reject H_0.

In the context of the problem our conclusion is:

The data provide sufficient evidence, at the 5% level of significance, to conclude that a majority of adults prefer the company's beverage to that of their competitor's.

6.4 Testing for a Difference in Two Population Means:

Independent samples are simple random samples from two distinct populations. To compare these random samples, both populations are normally distributed with the population means and standard deviations unknown unless the sample sizes are greater than 30. In that case, the populations need not be normally distributed.

The general steps of the hypothesis test in this case are actually the same as for one population mean. As expected, the details of the conditions for use of the test and the test statistic are unique to this test (but similar in many ways to what we have seen before).

Step 1: State the Null Hypothesis H_0 and Alternative Hypothesis H_a:

The **null hypothesis**, H_0, is again a statement of "no effect" or "no difference".

- **H_0: $\mu_1 - \mu_2 = 0$**, which is the same as **H_0: $\mu_1 = \mu_2$**

The **alternative hypothesis**, H_a, can be any one of the following.

- **H_a: $\mu_1 - \mu_2 < 0$**, which is the same as **H_a: $\mu_1 < \mu_2$**
- **H_a: $\mu_1 - \mu_2 > 0$**, which is the same as **H_a: $\mu_1 > \mu_2$**
- **H_a: $\mu_1 - \mu_2 \neq 0$**, which is the same as **H_a: $\mu_1 \neq \mu_2$**

Step 2: State the level of significance (α): Usually 0.05, 0.01, or 0.10

Step 3: Compute the value of the test statistic:

$$\mathbf{Z} = \frac{(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}{\sqrt{\frac{\boldsymbol{\sigma}_1^2}{\mathbf{n}_1} + \frac{\boldsymbol{\sigma}_2^2}{\mathbf{n}_2}}}$$

The term **$\mu_1 = \mu_2$** in the numerator disappears because we are assuming that **$\mu_1 = \mu_2$**, so **$\mu_1 = \mu_2$**= 0.

Note: In general, the population standard deviations are not known, and are estimated by the calculated values S_1 and S_2. In this case, the test statistic is defined as follows:

$$\mathbf{Z} = \frac{(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}{\sqrt{\frac{\mathbf{S}_1^2}{\mathbf{n}_1} + \frac{\mathbf{S}_2^2}{\mathbf{n}_2}}}$$

where

- S_1 and S_2, the sample standard deviations, are estimates of σ_1 and σ_2, respectively,
- σ_1 and σ_1 are the unknown population standard deviations,
- $\bar{X}_1$ and $\bar{X}_2$ are the sample means, and
- μ_1 and μ_2 are the population means.

Step 4: Determine Rejection Region:

One - Tailed (>): Reject H_0 if $Z > Z_{\frac{\alpha}{2}}$

Two - Tailed (≠): Reject H_0 if $Z > Z_{\frac{\alpha}{2}}$ or $Z < -Z_{\frac{\alpha}{2}}$

Critical z-values:

Test	α = 0.05		α = 0.01		α = 0.10	
One - tailed	(>) +1.645	(<) - 1.645	(>) +2.33	(<) - 2.33	(>) +1.28	(<) - 1.28
Two - tailed	$\pm$1.96		$\pm$ 2.58		$\pm$ 1.645	

Step 5: The decision is made as shown before in the **Figures 6.1**, **6.2** and **6.3**.

Step 6: As always, we state our conclusion in context, usually by referring to the alternative hypothesis.

Example 6.7

To Compare Customer satisfaction levels of two competing cable television companies, 174 customers of Company 1 and 355 customers of Company 2 were randomly selected and were asked to rate their cable companies on a five-point scale, with 1 being least satisfied and 5 most satisfied. The survey results are summarized in the following table:

	Company (1)	**Company (2)**
Sample Size	n_1 = 174	n_2 = 351
Sample mean	$\bar{x}_1$ = 3.51	$\bar{x}_2$ = 3.21
Sample Standard Deviation	S_1 = 0.51	S_2 = 0.52

Test at the 1% level of significance whether the data provide sufficient evidence to conclude that Company 1 has a higher mean satisfaction rating than does Company 2. Use the significance level 0.01.

Solution:

Step 1: If the mean satisfaction levels μ_1 and μ_2 are the same then $\mu_1 = \mu_2$, but we always express the null hypothesis in terms

of the difference between μ_1 and μ_2, hence H_0 is $\mu_1 - \mu_2 = 0$. To say that the mean customer satisfaction for Company 1 is higher than that for Company 2 means that $\mu_1 > \mu_2$, which in terms of their difference is $\mu_1 - \mu_2 > 0$. The test is therefore

$$\mathbf{H_0: \mu_1 - \mu_2 = 0 \quad vs \quad H_a: \mu_1 - \mu_2 > 0}$$

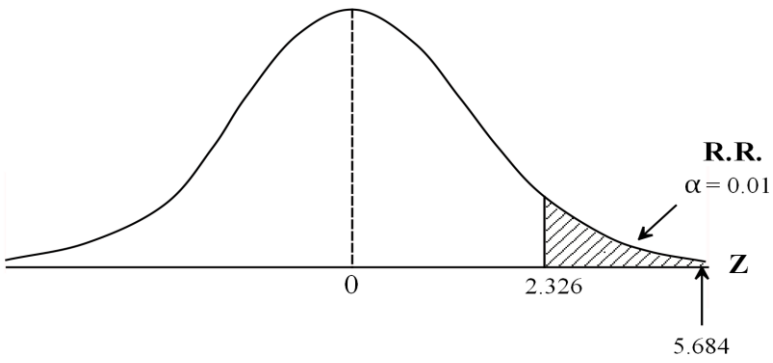
Step 2: The level of significance required is 0.01.

Step 3: Inserting the data into the formula for the test statistic gives

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

$$= \frac{(3.51 - 3.24)}{\sqrt{\frac{(0.51)^2}{174} + \frac{(0.52)^2}{355}}} = 5.684$$

Step 4: Since the symbol in H_a is ">" this is a right-tailed test, so there is a single critical value, $z_\alpha = 2.33$. The rejection region is $[2.33 , \infty]$ and shown in the following figure.



Step 5: As shown in this figure, the value of the test statistic falls in the rejection region. The decision is to reject H_0.

Step 6: In the context of the problem our conclusion is:

The data provide sufficient evidence, at the 0.01 level of significance, to conclude that the mean customer satisfaction for Company 1 is higher than that for Company 2.

Example 6.8

An interesting research question is the effect, if any, that different types of teaching formats have on the grade outcomes of students. To investigate this issue one sample of students' grades was taken from a hybrid class and another sample taken from a standard lecture format class. Both classes were for the same subject. The mean course grade in percent for the 35 hybrid students is 74 with a standard deviation of 16. The mean grades of the 40 students form the standard lecture class was 76 percent with a standard deviation of 9.

Test at 5% to see if there is any significant difference in the population mean grades between Class A and Class B.

Solution:

Step 1: We begin by noting that we have two groups, students from class A and students from class B. We also note that the random variable, what we are interested in, is students' grades, a continuous random variable. There is no presumption as to which format might lead to higher grades so the hypothesis is stated as a two-tailed test.

$$H_0: \mu_1 - \mu_2 = 0$$
$$H_a: \mu_1 - \mu_2 \neq 0$$

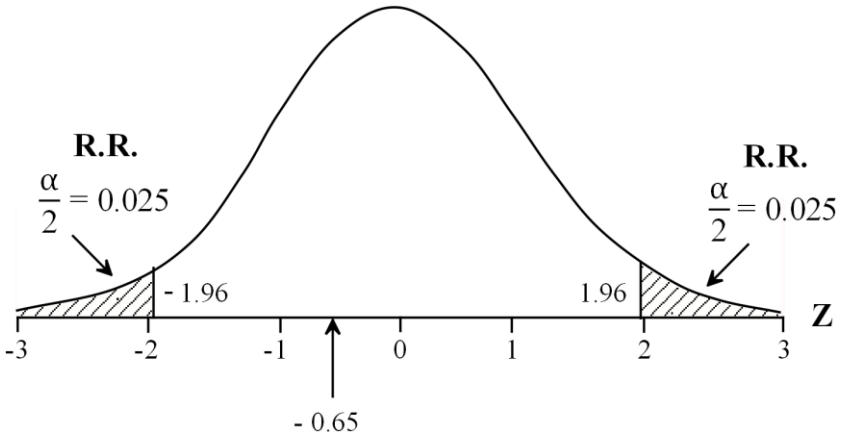
Step 2: The level of significance required is 0.05.

Step 3: As would virtually always be the case, we do not know the population variances of the two distributions and thus our test statistic is:

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

$$= \frac{(74 - 76)}{\sqrt{\frac{(16)^2}{35} + \frac{(9)^2}{40}}} = -0.65$$

Since the symbol in H_a is "≠" this is a two-tailed test, so there are two critical values, $\mathbf{Z}_{\frac{\alpha}{2}} = 1.96$. The rejection regions are $(-\infty , -0.65)$ and $[0.65 , \infty)$. The rejections regions are shown in the following figure.



Step 5: As shown in this figure, the value of the test statistic falls in the nonrejection region. The decision is not to reject H_0.

Step 6: Conclusion

In the context of the problem our conclusion is:

We cannot reject the null at α = 0.05. Therefore, evidence does not exist to prove that the grades in A and B classes differ.

Example 6.9

Following are data of weight loss for two groups (Diet and Exercise):

Weight Loss for Diet vs Exercisers:

	Sample Mean	Sample Standard Deviation	Sample Size
Diet Only	5.9 kg	4.1 kg	42
Exercise Only	4.1 kg	3.7 kg	47

Did exercisers lose less fat than dieters?

Solution:

Step 1. Determine the null and alternative hypotheses:
Null hypothesis: No difference in average fat lost in population for two methods. Population mean difference is zero.
Alternative hypothesis: Exercisers loss less fat than dieters.
Therefore:

H_0: $\mu_2 - \mu_1 = 0$
H_a: $\mu_2 - \mu_1 < 0$

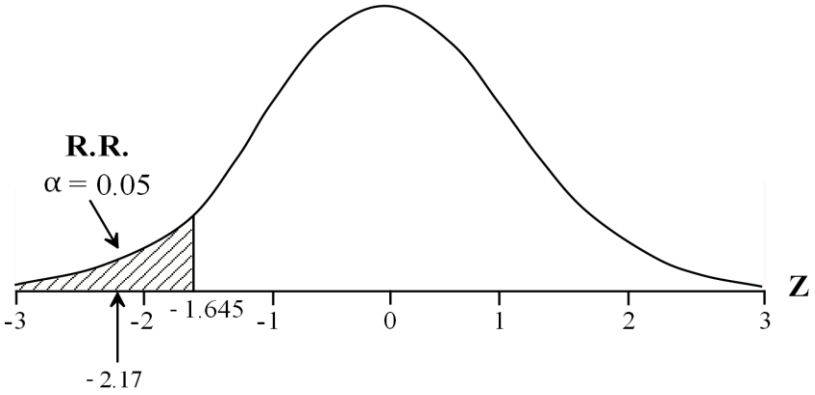
Step 2: The level of significance required is 0.05.

Step 3: As would virtually always be the case, we do not know the population variances of the two distributions and thus our test statistic is:

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

$$= \frac{(4.1 - 5.9)}{\sqrt{\frac{(3.7)^2}{47} + \frac{(4.1)^2}{42}}} = -2.17$$

Step 4: Since the symbol in H_a is "<" this is a left-tailed test, so there is one critical value, $z_\alpha = -1.645$. The rejection region is $(-\infty, -1.645)$. The rejection region is shown in the following figure.



Step 5: As shown in this figure, the value of the test statistic falls in the rejection region. The decision is to reject H_0.

Step 6: Conclusion

In the context of the problem our conclusion is:

We reject the null at $\alpha = 0.05$. Therefore, sample data provide sufficient evidence to conclude that exercisers lose less fat than dieters. That is, diet is more effective than exercise for reducing weight.

Note: The null and alternative hypotheses used in this example are the same as: $\mathbf{H_0: \mu_1 - \mu_2 = 0}$ **and** $\mathbf{H_a: \mu_1 - \mu_2 > 0}$

In this case, the test becomes a right - tailed test and will lead to the same result.

6.5 Testing for a Difference in Two Population Proportions:

When conducting a hypothesis test that compares two independent population proportions, the following characteristics should be present:

The test procedure, called the **two - proportion z-test**, is appropriate when the following conditions are met:

- The sampling method for each population is simple random sampling.
- The samples are independent.

- Each sample includes at least 5 successes and 5 failures.
- Each population is at least 20 times as big as its sample.

Comparing two proportions, like comparing two means, is common. If two estimated proportions are different, it may be due to a difference in the populations or it may be due to chance. A hypothesis test can help determine if a difference in the estimated proportions reflects a difference in the population proportions.

Skipping most of the details, the null hypothesis is the assumed condition that the proportions from both populations are equal, $\mathbf{H_0: P_1 - P_2 = 0}$, that is $\mathbf{H_0: P_1 = P_2}$, and the alternative hypothesis is one of the three conditions of non-equality.
When calculating the test statistic Z (notice we use the standard normal distribution), we are assuming that the two population proportions are the same, $P_1 = P_2 = \hat{P}$. Now if both Population 1 and Population 2 are the same in terms of the required proportion, they could be considered to be the "same" population. (Think about this a bit.) We define p̂ to be the **pooled** population proportion:

$$\hat{\mathbf{P}} = \frac{\mathbf{x_1 + x_2}}{\mathbf{n_1 + n_2}}$$

Substituting p̂ into the sample standard deviation expression gives:

$$\boldsymbol{\sigma_{\hat{p}}} = \sqrt{\frac{\hat{\mathbf{P}}(\mathbf{1} - \hat{\mathbf{P}})}{\mathbf{n_1}} + \frac{\hat{\mathbf{P}}(\mathbf{1} - \hat{\mathbf{P}})}{\mathbf{n_2}}} = \sqrt{\hat{\mathbf{P}}(\mathbf{1} - \hat{\mathbf{P}})}\sqrt{\frac{\mathbf{1}}{\mathbf{n_1}} + \frac{\mathbf{1}}{\mathbf{n_2}}}$$

The formula for the test statistic Z becomes:

$$\mathbf{Z} = \frac{\left(\hat{\mathbf{P}}_\mathbf{1} - \hat{\mathbf{P}}_\mathbf{2}\right) - (\mathbf{P_1} - \mathbf{P_2})}{\sqrt{\hat{\mathbf{P}}\left(\mathbf{1} - \hat{\mathbf{P}}\right)\left(\frac{\mathbf{1}}{\mathbf{n_1}} + \frac{\mathbf{1}}{\mathbf{n_2}}\right)}} = \frac{\left(\hat{\mathbf{P}}_\mathbf{1} - \hat{\mathbf{P}}_\mathbf{2}\right) - \mathbf{0}}{\sqrt{\hat{\mathbf{P}}\left(\mathbf{1} - \hat{\mathbf{P}}\right)\left(\frac{\mathbf{1}}{\mathbf{n_1}} + \frac{\mathbf{1}}{\mathbf{n_2}}\right)}}$$

$$= \frac{\left(\widehat{\mathbf{P}}_1 - \widehat{\mathbf{P}}_2\right)}{\sqrt{\widehat{\mathbf{P}}(1 - \widehat{\mathbf{P}})\left(\frac{1}{\mathbf{n}_1} + \frac{1}{\mathbf{n}_2}\right)}}$$

The term $P_1 - P_2$ in the numerator disappears because we are assuming that $P_1 = P_2$, so $P_1 - P_2 = 0$.

All other steps for the hypothesis test remain the same as discussed before.

The difference of two proportions follows an approximate normal distribution. Generally, the null hypothesis states that the two proportions are the same. That is, H_0: $P_1 = P_2$. To conduct the test, we use a pooled proportion, .

Hypothesis Testing Procedure:

Step1. State the Hypotheses:

Every hypothesis test requires the analyst to state a null hypothesis and an alternative hypothesis. The table below shows three sets of hypotheses. Each makes a statement about the difference *d* between two population proportions, P_1 and P_2. (In the table, the symbol ≠ means " not equal to ").

Null Hypothesis	Alternative Hypothesis	Number of Tails
$P_1 - P_2 = 0$	$P_1 - P_2 \neq 0$	2
$P_1 - P_2 = 0$	$P_1 - P_2 < 0$	1
$P_1 - P_2 = 0$	$P_1 - P_2 > 0$	1

The first set of hypotheses is an example of a two - tailed test, since an extreme value on either side of the sampling distribution would cause a researcher to reject the null hypothesis. The other two sets of hypotheses are one-tailed tests, since an extreme value on only one side of the sampling distribution would cause a researcher to reject the null hypothesis.

When the null hypothesis states that there is no difference between the two population proportions (i.e., $P_1 - P_2 = 0$), the null and alternative hypothesis for a two-tailed test are often stated in the following form:

H_0: $P_1 = P_2$

H_a: $P_1 \neq P_2$

Step2. Select the level of Significance:

The analysis plan describes how to use sample data to accept or reject the null hypothesis. It should specify the significance level. Often, researchers choose significance levels equal to 0.01, 0.05, or 0.10; but any value between 0 and 1 can be used.

Step 3. Compute the Value of the Test Statistic:

Using sample data, complete the following computations to find the test statistic.

- **Pooled sample proportion.** Since the null hypothesis states that $P_1 = P_2$, we use a pooled sample proportion (p) to compute the standard error of the sampling distribution.

$$\widehat{\mathbf{P}} = \frac{\mathbf{x_1} + \mathbf{x_2}}{\mathbf{n_1} + \mathbf{n_2}}$$

  where p_1 is the sample proportion from population 1, p_2 is the sample proportion from population 2, n_1 is the size of sample 1, and n_2 is the size of sample 2.

- **Standard error.** Compute the standard error (SE) of the sampling distribution difference between two proportions.

$$\textbf{Standard Error (SE)} = \sqrt{\widehat{\mathbf{P}}(\mathbf{1} - \widehat{\mathbf{P}})\left(\frac{\mathbf{1}}{\mathbf{n_1}} + \frac{\mathbf{1}}{\mathbf{n_2}}\right)}$$

  where $\hat{p}$ is the pooled sample proportion, n_1 is the size of sample 1, and n_2 is the size of sample 2.

- **Test statistic.** The test statistic is a z-score (z) defined by the following equation:

$$Z = \frac{\left(\hat{P}_1 - \hat{P}_2\right)}{\sqrt{\hat{P}(1-\hat{P})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where p_1 is the proportion from sample 1, p_2 is the proportion from sample 2, and SE is the standard error of the sampling distribution.

Step 4. Determine the Rejection Region:
Determine Rejection Region:

One - Tailed (>): Reject H_0 if $Z > Z_{\frac{\alpha}{2}}$

Two - Tailed (≠): Reject H_0 if $Z > Z_{\frac{\alpha}{2}}$ or $Z < -Z_{\frac{\alpha}{2}}$

The critical values used for different hypotheses are as given in **Table 6.1**.

Step 5. Make a Decision by determining if the calculated test statistic is in the rejection region or not (Reject or do not reject H_0).

Step 6. Interpreting Results (Conclusion):
Interpret the result of the hypothesis test. That is, conclusion is written in terms of the original problem.

Following are two sample problems illustrate how to conduct a hypothesis test for the difference between two proportions. **Example 6.10** involves a two - tailed test; **Example 6.11**, a one-tailed test.

Example 6.10

Suppose a Pharmaceutical Company develops a new drug, designed to prevent colds. The company states that the drug is

equally effective for men and women. To test this claim, they choose a simple random sample of 100 women and 200 men.
At the end of the study, 38% of the women caught a cold; and 51% of the men caught a cold. Based on these findings, can we reject the company's claim that the drug is equally effective for men and women? Use a 0.05 level of significance.

Solution:

We work through those steps below:

Step 1. State the Hypotheses:

The first step is to state the null hypothesis and an alternative hypothesis.

Null hypothesis: H_0: $P_1 - P_2 = 0$ which is equivalent to

H_0: $P_1 = P_2$

Alternative hypothesis: H_a: $P_1 - P_2 \neq 0$ Which is equivalent to

H_a: $P_1 \neq P_2$

Step 2. Select the Level of Significance:

The level of significance given is 0.05.

Step 3. Compute the Value of the Test Statistic:

Using sample data, the following computations are required to find the value of the test statistic.

- $\widehat{P} = \frac{(n_1\widehat{P}_1 + n_2\widehat{P}_2)}{(n_1+n_2)}$

$= \frac{[100 \times 0.38 + 200 \times 0.51]}{(100 + 200)}$

$= 140 \div 300 = 0.467$

- Standard Error (SE) $= \sqrt{\widehat{P}(1-\widehat{P})\left[\left(\frac{1}{n_1}\right) + \left(\frac{1}{n_2}\right)\right]}$

$= \sqrt{0.467 \times 0.533 \times \left[\left(\frac{1}{100}\right) + \left(\frac{1}{200}\right)\right]}$

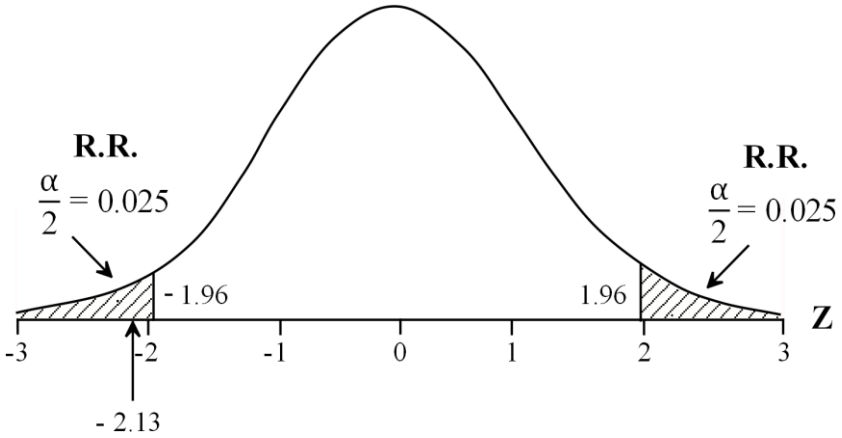
$= \sqrt{0.003733} = 0.061$

- The value of the test statistic is given by:

$$Z = \frac{(P_1 - P_2)}{SE} = \frac{(0.38 - 0.51)}{0.061} = -2.13$$

Step 4. Determine the Rejection Region:

Since the symbol in H_a is " $\neq$" this is a two - tailed test, so there are two critical values, $Z_{\frac{\alpha}{2}} = -1.96$ and $z_\alpha = 1.96$. The two rejection regions are $(-\infty , -1.96)$ and $[1.96 , \infty)$. The rejection regions are shown in the following figure:



Step 5: Make a Decision

As shown in this figure, the value of the test statistic falls in the rejection region. The decision is to reject H_0.

Step 6: Conclusion

In the context of the problem our conclusion is:

The data provide sufficient evidence, at the 0.05 level of significance, to conclude that the company's claim that the drug is equally effective for men and women is true.

Example 6.11

For the data of **Example 6.10,** suppose the company claims that the drug is more effective for women than for men.

Investigate the validity of this claim using a 0.01 level of significance.

Solution:

Step 1. State the Hypotheses:

In this case, the null hypothesis and an alternative hypothesis are:

Null hypothesis: H_0: $P_1 - P_2 = 0$ which is equivalent to

H_0: $P_1 = P_2$

Alternative hypothesis: H_a: $P_1 - P_2 < 0$ Which is equivalent to

H_a: $P_1 < P_2$

Step 2. Select the Level of Significance:

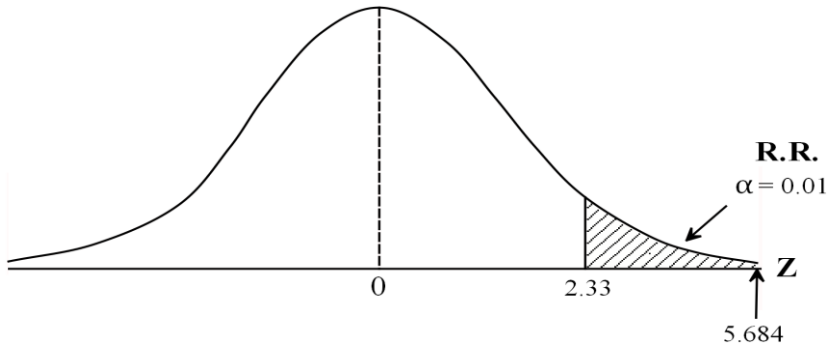
The level of significance given is 0.01.

Step 3. Compute the Value of the Test Statistic:

The value of the test statistic is – 2.13 (Example 6.10).

Step 4. Determine the Rejection Region:

Since the symbol in H_a is "< " this is a left - tailed test, so there is one critical value, z_α = - 2.33. The rejection region is (-∞ , -2.33). The rejection regions is shown in the following figure:



Step 5: Make a Decision

As shown in this figure, the value of the test statistic falls in the nonrejection region. The decision is:do not reject H_0.

Step 6: Conclusion

In the context of the problem our conclusion is:

The data do not provide sufficient evidence, at the 0.01 level of significance, to conclude that the company's claim that the drug is more effective for women than for men is true.

Exercises for Chapter 5 and Chapter 6

(Exam Questions)

Q1. If it is assumed that the heights of men are normally distributed, with a standard deviation of 2.5 inches, how large sample should be taken to be 95% sure that the sample mean does not differ from the true population mean by more than 0.5 inches in absolute value?

Q2. How large a sample must one take to be 99% confident that the estimate is within 0.05 of the true proportion of women over 55 who are widows? If**:**

(a) A recent study indicates that 20% of women over 55 in the study were widows.

(b) No information available about the population proportion.

(c) Comment on the results of parts (a) and (b).

Q3. In a survey of 250 voters prior to an election, 48% indicated that they would vote for candidate A.

1. What is the point estimate of the population proportion of voters who support candidate A?
2. Make a 95% confidence interval for the population proportion. Interpret this interval.
3. What is the maximum error of estimate for part (b).
4. How can you reduce the width of the confidence interval? describe all possible alternatives, which alternative do you think is the best and why?
5. Does the confidence interval constructed in part (c) include the value p = 0.45? What might you conclude?
   It is claimed that less than half of the voters will support candidate A, investigate the validity of this claim using $\alpha = 0.01$. **(Exam 1998)**↓

Q4. The following table shows the distribution of monthly wages of samples of employees from two companies.

Company	Sample Size	Mean	Standard Deviation
A	40	L.E. 300	20
B	50	L.E. 280	25

(a) Construct a 99% confidence interval for μ_1. Interpret this interval.

(b) What are the **MINITAB** commands required for constructing the confidence interval of part (a)? Present the **MINITAB** output for these commands using the results of part (a)

(c) Find the point estimate of $\mu_1 - \mu_2$ where μ_1 is the mean monthly wages for all employees in company A, and μ_2 is the mean monthly wages for all employees in company B.

(d) Construct a 99% confidence interval for $\mu_1 - \mu_2$.

(e) Assuming that the two sample observations are available, what are the **MINITAB** commands that are required for determining the confidence interval of part (d)?

(f) Based on the results of part (d), test at the 1% significance level if the mean monthly wages of all employees in the two companies are different.

(g) Would your results in part (f) change if you use the 5% confidence level and why? (No calculations required).

(h) The personnel manager in company A claims that the wages in his company are significantly higher than those in company B. Do you think the sample information supports this claim? Use $\alpha = 0.05$.

(i) Assuming that the two samples observations are available, state the **MINITAB** commands required to perform the test of part (h). **(Exam 1998)**

Q5. A study is conducted to compare the proportion of satisfied new-car buyers for domestic and foreign models. In a sample of **600** buyers of new domestic cars, **480** reported they were highly satisfied with their cars. In a sample of **450** buyers of new foreign cars, **369** indicated high satisfaction.

Let $\mathbf{P_1}$ is the proportion of satisfied buyers for domestic cars.

$\mathbf{P_2}$ is the proportion of satisfied buyers for foreign cars.

(1) Provide a point estimate for $\mathbf{(P_2 - P_1)}$.

(2) Determine a **95%** confidence interval for the **difference** between the two population proportions $\mathbf{(P_2 - P_1)}$. Interpret your interval.

(3) Based on the results of **Part (2)**, do you think that there is a significant difference in the proportions for the two models in the general populations? Explain.

(4) If the buyers of foreign models claim that they are more satisfied with their cars than those of domestic models. Using the 1% significance level, investigate the validity of this claim.

Q6. An experiment was conducted to compare two diets **A** and **B** designed for weight reduction. Two groups, each of size **50** overweight dieters, were randomly selected. One group was placed on diet A and the other on diet B, and their weight losses were recorded over a 30 - day period. The means and standard deviations of the weight-loss measurements for the two groups are shown below**:**

	Diet A	**Diet B**
Mean	14.5	13.4
Standard Deviation	2.6	1.9

(1) Find a **99%** confidence interval for (μ_A). Explain your results.

(2) Make the following tests: (**Exam 1999)** ↓

(a) $H_0: \mu_B = 14$ **Vs** $H_1: \mu_B < 14$

(b) $H_0: \mu_B = 14$ **Vs** $H_1: \mu_B \neq 14$

(3) Determine a **95%** confidence for the difference in mean weights loss for the two diets. Interpret your confidence interval.

(4) Do the interval in **Part (c)** contain the value **zero**? Why would this is of interest to the researcher?

(5) Can you conclude that Diet (A) is more effective than Diet (B) for reducing Weight? Use the 1% significance level.

(6) Assuming that the observations of the two samples are available, use **MINITAB** to answer the **Parts (1)** and **(2)**.

Q7. A marketing manager is in the process of deciding whether to introduce a new product. He has concluded that he needs to perform a market survey in which he asks a random sample of people whether they will buy the product. Find the most **conservative sample size** he should take if he wants to estimate the proportion of people who will by the product to within **0.03** with **99%** confidence.

Q8. The operation manager of a large production plant would like to estimate the average amount of time a worker takes to estimate a new electronic component. How large a sample of workers should he take if he wishes to estimate the mean assembly time to within **20** seconds using **99%** confidence level if**:**

(1) From earlier study, the manager knows that the standard deviation of amount of time taken by workers to assemble a similar device is **2.5** minutes.

(2) After observing a number of workers assembling similar devices, he noted that the shortest time taken was **10** minutes, while the longest time taken was **22** minutes.

(Exam 1999)

Q9. A medical researcher wishes to see whether the pulse rates of smokers are higher than the pulse rates of nonsmokers. Samples of 100 smokers and 100 nonsmokers are selected. The results are shown here.

Smokers	**Nonsmokers**
$\bar{x}_1 = 90$	$\bar{x}_2 = 88$
$s_1 = 5$	$s_2 = 6$
$n_1 = 100$	$n_2 = 100$

(a) Provide a point estimate for μ_1, where μ_1 is the mean pulse rate of all smokers. What is the margin of error associated with this point estimate?

(b) Construct a 99% confidence interval for the mean pulse rate of all smokers (μ_1). Explain.

(c) Using the results obtained in **Part (b)**, can you say that the mean pulse rate of all smokers equals 91?

(d) What is the maximum error of estimate for **Part (b)**?

(e) Determine a 95% confidence interval for the difference between the two population means ($\mu_1 - \mu_2$). Explain.

(f) Can the researcher conclude, at $\alpha = 0.05$, that smokers have higher pulse rates than nonsmokers? What is the type 1 error in this case? Explain in words.

(g) Find the p-value for the test in **Part (f)**.

(h) What will your decision be in part (f) if the probability of making a type I error is zero? Explain.

(i) If $\alpha = 0.01$, based on the p-value calculated in part (g), would you reject the null hypothesis? Explain.

(j) Suppose the observations of the two samples are available use **MINITAB** to answer **Parts (b)**, **(e)**, and **(f)**.

(Exam 2000) ↓

Q10. A company wanted to know if attending a course on "how to be a successful salesperson" can increase the average sales of its employees. The company sent six of its salespersons to attend this course. The following table gives the one-week sales of these salespersons before and after they attended the course.

Before	12	18	25	9	14	16
After	18	24	24	14	19	20

(a) Construct a 99% confidence interval for the mean μ_d of the population paired differences where a paired difference is equal to the weekly sales after attending this course minus the weekly sales before attending this course.

(b) Using the 1% significance level, can you conclude that the mean weekly sales for all salespersons increase as a result of attending this course? Assume that the population of paired differences has a normal distribution.

(c) Use **MINITAB** to answer part (a).

Q11. Suppose the president wants an estimate of the proportion of the population who support his current policy toward Israel. The president wants the estimate to be within 0.04 of the true proportion. Assume a 95% confidence level. A recent study estimated the proportion supporting current policy to be 0.8.

(a) How large a sample is required?

(b) How large would the sample have to be if the estimate of the recent study were not available?

(c) How can you explain the difference between the two sizes of the sample you reached in parts (a) and (b)?

(Exam 2000)

Q12. A sample of 200 male registered voters showed that 40% of them voted in the last election. Another sample of 100 female registered voters showed that 38% of them voted in the same election.

(a) Construct a 95% confidence interval for the difference between the proportion of all male and all female registered voters who voted in the last election. Explain.

(b) Test at the 1% significance level if the proportion of all male voters who voted in the last election is greater than 36%. Explain.

Q13. The operations manager of a large production plant would like to estimate the average amount of time a worker takes to assemble a new electronic component. After observing a number of workers assembling similar devices, he noted that the shortest time taken was 10 minutes, while the longest time was 22 minutes. How large a sample of workers should he take if he wishes to estimate the mean assembly time to within 54 seconds? Assume that the confidence level is to be 99%.

Q14. An experiment was planned to compare the mean time (in days) required to recover from a common cold for persons given a daily dose of 4 milligrams of vitamin C versus those who were not given a vitamin supplement. Suppose that 50 adults were randomly selected for each treatment category and that the mean recovery times and standard deviations for the two groups were as follows**:**

(Exam 2001) ↓

	Treatment	
	No Vitamin Supplement	**4 mg Vitamin C**
Sample size	50	50
Sample mean	5	4
Sample standard deviation	3	1
Population mean	μ_1	μ_2

Suppose your research objective is to show that the use of vitamin C reduces the mean time required to recover from a common cold and its complications.

(a) Give the null and alternative hypotheses for the test.

(b) Conduct the statistical test of the null hypothesis in Part (a) and state your conclusion. Test using $\alpha = 0.05$.

(c) Construct a 99% confidence interval for μ_1.

(d) Supposing the observations are available for the first sample, use **MINITAB** to construct the confidence interval in Part (c).

(Exam 2001)

Q15. A statistician claims that the mean score on a standardized test of students who major in accounting is greater than that of students who major in economics. The results of the test, given to **50** students in each group, are shown here.

Accounting	**Economics**
$\bar{x}_1 = 118$	$\bar{x}_2 = 115$
$s_1 = 15$	$s_2 = 12$
$n_1 = 50$	$n_2 = 50$

(a) What is the point estimate of the difference between the two population means **$(\mu_1 - \mu_2)$?**

(b) Construct and interpret a **95%** confidence interval for **$(\mu_1 - \mu_2)$**.

(Exam 2002) ↓

(c) What is the **maximum error** of estimate for **Part (b)?**

(d) Make a **99%** confidence interval for μ_2.

(e) Testing at the **1%** significance level, can you conclude that the mean score of all students who major in economics is different from **110?**

(f) Is there sufficient evidence in the samples to support the statistician's claim at $\alpha = 0.05$**?**

Q16. In a school, it was found that a preliminary random sample of **40** pupils contains **five** who are left-handed.

(a) Construct a **95%** confidence interval for the proportion of pupils in the school who are left-handed.

(b) How large the sample have been to reduce the width of this confidence interval to **0.1?**

(Exam 2002)

Q17. The dean of students wants to find out if there is any significant difference in the mathematical ability of male and female students as determined by their achievement scores in a basic skills test in mathematics. A random sample of **150** male students and **100** female students was selected from all the students who took the test. Their results are summarized as follows**:**

	Male Students	**Female Students**
Mean	**85**	**80**
Standard Deviation	**16**	**12**

(a) What is the point estimate of the difference between the two populations means**?**

(b) Construct a **95%** confidence interval for the mean score of all **male** students**.**

(Exam 2003 Qena) ↓

(c) What is the maximum error of estimate for **Part (b)?**

(d) Suppose the confidence interval obtained in **Part (b)** is too wide. How can the width of this interval be reduced**?** Discuss all possible alternatives. Which of these alternatives is the best**?**

(e) Make a **99%** confidence interval for the difference between the two population means.

(f) Using the **5%** significance level, can you conclude that the mathematical ability of male students is better than that of female students**?**

Q18. The accounting department of a major bank has noticed an increase in the proportion of delinquent customers who don't pay back their loan on time. To find an estimate for current delinquency, the manager plans to investigate the payment records of a sample of customers. The manager does not want the error of his estimate to exceed **3%**. At a confidence level of **95%**, what size sample should be selected if

(a) A study conducted several years ago revealed that the proportion of delinquent customers was **20%**.

(b) No such previous estimate is available

(Exam 2003 Qena)

Q19. A random sample of **150** industrial firms of **Type A** shows that **27%** of them spend more than 3% of their total sales on advertising. A similar independent sample of **100** industrial firms of **Type B** shows that **20%** of them spend more than 3% of their total sales on advertisement.

(a) What is the point estimate of the difference between the two population proportions $\mathbf{(P_A - P_B)}$.

(b) Construct a **99%** confidence interval for the proportion $\mathbf{P_B}$.

(Exam 2003 Sohag) ↓

(c) Assuming that the sample proportion and confidence level remain the same as in **(b)**, how large should the sample have been to reduce the width of the confidence interval constructed in Part (b) by **0.02?**

(d) Construct a **95%** confidence interval for the difference between the two population proportions $\mathbf{(P_A - P_B)}$.

(e) Using the **5%** significance level, can you conclude that the **same** proportion of **Type A** and **Type B** industrial firms spend more than 3% of their total sales on advertising.

(f) Based on the results obtained in Part (d), how to verify your results in Part (e).

Q20. An efficiency expert in a large assembly plant for laptop computers is interested to determine the average time it takes a worker to assemble a laptop computer with available parts.

(a) How large a sample will he need to be **95%** confident that his sample mean will not differ from the true mean by more than **10** minutes**?** Similar studies conducted before have established a standard deviation of **40** minutes.

(b) How would the sample size change if the confidence level is changed to **99%?** Other conditions remain the same as in **part (a)** above.

(c) How can you explain the difference between the two sizes of the sample you reached in Parts (a) and (b)**?**

(Exam 2003 Sohag)

Q21. A consulting agency was asked by a large insurance company to investigate if business majors were better salespersons. A sample of **40** salespersons with a business degree showed that they sold an average of **10** insurance policies per week with a standard deviation of **1.8** policies. Another sample of **45** salespersons with a degree other than

(Exam 2004) ↓

business showed that they sold an average of **8.5** insurance policies per week with a standard deviation of **1.35** policies.

(a) Construct a **99%** confidence interval for the difference between the two population means **($\mu_1 - \mu_2$)**.

(b) Make and interpret a **95%** confidence interval for **μ_2.**

(c) Test at the **5%** significance level if **μ_1** is different from **10.5**.

(d) Using the **1%** significance level can you conclude that persons with a business degree are better salespersons than those who have a degree in another area**?**

(e) Assuming that the observations of each sample are available, use **MINITAB** to answer Parts **(a)**, **(b)**, **(c)**, and **(d)**.

Q22. An estimate is required of the proportion of a large number of consumers who are likely to purchase a particular brand of butter. Determine the sample size that should be taken so that the **99%** confidence interval for the population proportion has a maximum error of **0.02** in each of the following situations**:**

(a) The population proportion is known to be **0.6**.

(b) Nothing is known about the value of the population proportion.

(c) Comment on the results of **Parts (a)** and **(b)**.

Q23. The 1977 Statistical Abstract of the United States reports the percentage of people 18 years of age and older who smoke. Assume that a study is being designed to collect new data on smokers and nonsmokers. The best preliminary estimate of the population proportion who smoke is **30%**.

(1) How large a sample should be taken to estimate the proportion of smokers in the population with a maximum error of **0.02?** Use **95%** confidence level.

(Exam 2004) ↓

(2) If the desired maximum error is fixed, what happens to the sample size as the confidence level is increased**?** Explain your reasoning without performing any calculations.

(3) Assume that the study uses your sample size recommendation in **Part (1)** and finds **580** smokers.

(a) What is point estimate of the proportion of smokers in the population**?**

(b) What is the **95%** confidence interval estimate for the proportion of smokers in the population**?**

(c) Is the result of **Part (b)** in agreement with the statement that **30%** of population are smokers**?** Explain.

(d) Using α **= 0.01**, does the sample provide sufficient evidence to conclude that the proportion of smokers in the population is **less than 0.3? Hint:** The sample size (n) is as obtained in **Part (1)**, and the sample proportion is $\widehat{\mathbf{P}} = \frac{\mathbf{580}}{\mathbf{n}}$).

Q24. A firm is studying the delivery times of two raw material suppliers. The firm is basically satisfied with supplier **A** and is prepared to stay with that supplier if the mean delivery time is the same as or less than that of supplier **B**. However, if the firm finds that the mean delivery time of supplier **B** is less than that of supplier **A**, it will begin making raw material purchases from supplier **B**.

Assume that independent samples show the following delivery time characteristics for the two suppliers.

	Supplier A	**Supplier B**
Size	$n_1 = 50$	$n_2 = 30$
Mean	$\bar{x}_1 = 14$ **days**	$\bar{x}_2 = 12.5$ **days**
Standard Deviation	$s_1 = 3$ **days**	$s_2 = 2$ **days**

(Exam 2004) ↓

(1) (a) What are the null and alternative hypotheses for this situation**?**

(b) With α **= 0.05**, what is your conclusion for the hypotheses from **Part (a)?**

(c) What action do you recommend in terms of supplier selection**?**

(d) Would your results in **Part (b)** change if you use the **1%** significance level**?**

(2) Is there sufficient evidence in the first sample that the mean delivery time of **supplier A** differs significantly from **15** days**?** Use a significance level of α **= 0.05**.

(3) Provide a **95%** confidence interval for the difference between the two population means **(μ_1 - μ_2)**.

(4) What is the maximum error of estimate for **Part (3)?**

(Exam 2004)

Q25. A large automobile insurance company selected samples of single and married male policyholders and recorded the number who had made an insurance claim over the preceding three - years period.

Single Policyholders	**Married Policyholders**
n_1 = 400	**n_2 = 900**
Number making claims = **76**	Number making claims = **90**

(1) Provide a **95%** confidence interval for the population proportion **P_1**.

(2) What is the point estimate for P1 – P2?

(3) What is the 99% confidence interval for the difference between the proportions for the two populations?

(4) Does the first sample provide sufficient evidence to conclude that the proportion of the single male policyholders in the population is less than **0.2**. Use α **= 0.05**. **(Exam 2005)** ↓

(5) Does the first sample provide sufficient evidence to conclude that the proportion of the single male policyholders in the population is less than **0.2**. Use α **= 0.05**.

(5) Using the **1%** significance level, determine whether the claim rates differ between single and married male policyholders.

Q26. A sample survey is to be conducted to determine the mean family income in an area. The question is, how many families should be should be sampled**?** In order to get more information about the area, a small pilot survey was conducted, and the standard deviation of the sample was computed to be \$**500**. The sponsor of the survey wants to use the **95%** degree of confidence. The desired margin of error is \$**100**.

(1) How many families should be interviewed**?**

(2) Suppose the sponsor wishes to reduce the margin of error by \$**20**, how large a sample is required**?**

(3) Compare between your results of **Parts (1)** and **(2)?** Explain.

Q27. The production manager of a manufacturing company produces certain electrical components believes that the life time of a particular kind of components produced by **Process A** is greater than the mean life time of the same kind produced by **Process B**. A random sample was selected from that component produced by each process and examined. The following results were obtained.

Process	Sample Size	Mean Life Time	Standard Deviation
A	40	250 hours	25 hours
B	50	235 hours	20 hours

(1) Using the **1%** significance level, is there evidence that the manager is justified in his belief?

(Exam 2005) ↓

(2) Develop a **99%** confidence interval for the difference between the two population means ($\mu_A - \mu_B$).

(3) Construct a **95%** confidence interval for the mean life time of the component produced by **Process B** (μ_B).

(4) Is it possible that the population mean μ_B is **238?** Explain.

(5) It is claimed that the mean life time of the component produced by **Process B** differs from **238** hours. Investigate the validity of this claim using α **= 0.05**.

(6) Do the results of **Part (5)** agree with those obtained for **Parts (3)** and **(4)? Explain**.

(Exam 2005)

Q28. A company manufacturing computers establishes an aptitude test for potential programmers. During the first three months, **100** candidates take the test. The company is interested in discovering whether candidates with a mathematical background show greater aptitude than those without this background. The results of the tests are as follows**:**

	Candidates	
	With Mathematics	**Without Mathematics**
Number of Candidates	40	60
Mean	85	82
Standard Deviation	10	12

(1) What is the point estimate of **($\mu_2 - \mu_1$)?** Where μ_1 is the mean score of all candidates with a mathematical background and μ_2 is the mean score of all candidates without mathematical background.

(Exam 2006) ↓

(2) Make a **99%** confidence interval for the difference between the two population **(μ_1 - μ_2)**, and explain briefly what this means.

(3) What is the maximum error for the estimate of **Part (2)?**

(4) Do the samples provide sufficient evidence to conclude that the mean score of all candidate with a mathematical background show greater aptitude than those without this background**?** Use α **= 0.05**.

(5) Is there any evidence, at the 1% significance level, that these results indicate that the mean score of candidates without mathematical background is different from 84?

Q29. A manufacturer of alkaline batteries claims that at most **3%** of the produced batteries is defective. Suppose each of **500** randomly selected batteries is tested, and **18** defective batteries are found.

(1) Provide a **95%** confidence interval for the population proportion **P**.

(2) At α **= 0.01**, investigate the validity of the manufacturer's claim.

Q30. An estimate is required of the proportion of large number of consumers who are likely to purchase a particular brand of butter. What sample size would have to be taken in each of the following situations**:**

(1) The population proportion is known to be **0.8**, and it is required to be **99%** confident that the difference between the sample proportion and the population proportion is **0.024**.

(2) Nothing is known about the value of the population proportion and it is required to estimate the population proportion with a maximum error of **0.03** and a confidence level of **99%**.

(Exam 2006) ↓

(3) Compare between your results of **Parts (1)** and **(2)**. Explain.

Q31. A random sample of **500** fish is taken from a lake, marked, and returned to the lake. After a suitable interval a second sample of **500** is taken and **25** of these are found to be marked.

(1) Find the **point estimate** of the proportion of marked fish in the lake.

(2) Estimate the number of fish in the lake.

(3) Obtain a **95%** confidence interval for the number of fish in the lake.

Exam 2006

Q32. A research firm conducted a survey to determine the mean amount smokers spend on cigarettes during a week. A sample of 64 smokers revealed that the mean **($\bar{x}$)** is L.E. 20 and standard
deviation (s) is L.E. 4. Assume that the sample was drawn from a normal population.

Let **μ_A** is the **mean** price for **Brand (A)** and **μ_B** is the **mean** price for **Brand (B)**.

(1) What is the point estimate of the population mean **(μ)**

(2) Using the 95% level of significance, determine the confidence interval for the population mean **(μ)**.

(3) Repeat Part (2) assuming that you know that the population standard deviation is $\sigma = 4$.

(4) Explain why the interval estimate produced in **Part (3)** is **narrower** than the one determined in **Part (2)**.

(5) It is claimed that the mean amount smokers in the population spend on cigarette during a week is greater than L.E. 18. Investigate the validity of this claim using **$\alpha = 0.05$**. State your conclusion in the language of this issue. Assume that **$\sigma = 4$**.

(Exam 2007) ↓

Q33. A statistician estimates the **95%** confidence interval for the mean of a normally distributed population as **140.2** to **159.8**. What is the **99%** confidence interval?

Q34. The technical institute (A) claims "94% of our graduates get jobs" Assume that the result is based on a random sample of 100 graduates of the program. Suppose that an independent random sample of 125 graduates of a competing technical institute (B) reveals that 92% of these graduates got jobs.

Let $\mathbf{P_1}$ and $\mathbf{P_2}$ be the **proportions** of the graduates who got jobs in general in the institutes **A** and **B** respectively.

(1) Provide a **99%** confidence interval for the **difference** between the two population proportions $\mathbf{(P_A - P_B)}$.

(2) Is there evidence to conclude that one institute is more successful than the other in placing its graduates? Use $\alpha = 0.01$.

(3) Would your results in **Part (2) change** if you use $\boldsymbol{\alpha = 0.05}$? Explain.

(4) What happens to the maximum error as the confidence level in **Part (1)** is decreased from 99% to 95%? Hint: Do not calculate the lower and upper limits for the 95% confidence interval for the difference between the two population proportions $\mathbf{(P_A - P_B)}$.

Q35. The proportion of public accountants who have changed companies within the last 3 years is to be estimated within 0.03. The 95% level of confidence is to be used. A study conducted several years ago revealed that the proportion of public accountants changed companies within 3 years was 0.21.

(1) To update this study, the files of how many public accountants should be studied?

(2) How many public accountants should be contacted if no previous estimates of the population proportion are available?

(Exam 2007) ↓

Q36. A student conducted a study and reported that the 95% confidence interval for the population mean ranged from 46 to 54. He was sure that the mean of the sample was 50, that the standard deviation of the sample was 16, and that the sample was at least 30, but could not remember the exact number. Can you help him out?

(Exam 2007)

Q37. An experiment has been conducted to compare the productivity of **two** machines. **Machine 1** was observed for **40** hours and **Machine 2** for **50** hours. The average productivity of items produced per hour and the standard deviation for each machine are recorded below**:**

	Machine 1	**Machine 2**
Average	**61.4**	**59.5**
Standard Deviation	**3.1**	**2.8**

(1) What is the **point estimate** of $\mu_1 - \mu_2$**?** Where μ_1 and μ_2 are the average productivity of items produced per hour for **Machine 1** and **Machine 2**, respectively.

(2) Develop a **95%** confidence interval for $\mu_2 - \mu_1$.

(3) What is the **maximum error** for the estimate of **Part (2)?**

(4) Do the samples provide sufficient evidence to conclude that productivity on **Machine 1** is **better than** productivity on **Machine 2?** Use a significance level equal to **0.01**.

(5) Comment on whether it would be possible to reach a different conclusion for **Part (4)** using α **= 0.05** instead of α **= 0.01**. Explain.
No calculations required.

Q38. An efficiency expert in a large assembly plant for laptop computers is interested to estimate the average time it takes a worker to assemble a laptop computer with available parts.

(Exam 2008) ↓

(1) How large a sample will he need to be **95%** confident that his sample mean will not differ from the true mean by more than **15** minutes if**:**

(a) Similar studies conducted before have established a standard deviation of **50** minutes.

(b) After observing a number of workers assembling similar laptop, he noted that the **shortest** time taken was **2** hours, while the **longest** time taken was **5** hours.

(2) How can you explain the difference between the two sizes of the sample you reached in Parts **(a)** and **(b)?**

Q39. An insurance company believes that **smokers** have **higher incidence** of heart disease than **non-smokers** in men over 50 years of age. Accordingly, it is considering to offer discounts on its life insurance policies to **non-smokers**. However, before the decision can be made, an analysis is undertaken to justify its claim that the Smokers are at a higher risk of heart disease than non-smokers. The company randomly selected **200** men which included **80 smokers** and **120 nonsmokers**. The survey indicated that **18 smokers** suffered from heart disease and **15 non-smokers** suffered from heart disease.

Let $\mathbf{P_1}$ = **Proportion** of male **smokers** over 50 years of age who suffer from heart disease, and

$\mathbf{P_2}$ = **Proportion** of male **non-smokers** over 50 years of age who suffer from heart disease.

(1) At **5%** level of significance, can you justify the claim of the insurance company that the **non-smokers** have a **lower** incidence of heart disease than **smokers?**

(Exam 2008) ↓

(2) Provide a 99% confidence interval for the proportion of male smokers over 50 years of age who suffered from heart disease (P1). Explain.

(3) Using 1% significance level, can you conclude that the proportion of male smokers over 50 years of age who suffered from heart disease (P1) is different from 0.24?

(4) Are the results of the hypothesis test (Part 3) consistent with the confidence interval you produced in Part (2)? If so, discuss why; if not, discuss why not.

Q40. The accounting department of a major bank has noticed an increase in the proportion of delinquent customers who do not pay back their loan on time. To find an estimate for current delinquency, the manager plans to investigate the payment records of a sample of customers. The manager does not want the error of his estimate to exceed **2%**. At a confidence level of **99%**, what sample size should be selected if:

(1) A study conducted several years ago revealed that the proportion of delinquent customers was **25%**.

(2) Assume that the manager has **no knowledge** what proportion might be and wants to make sure he has a large enough sample size to meet his need**.**

(Exam 2008)

Q41. A motor-car manufacturer purchases gear assemblies from a sub-contractor who undertakes to ensure that not more than **5%** of his supplies will be defective. In order to provide a check on the quality of incoming supplies, a random sample of **200** assemblies is selected of which **18** are found to be defective.

(1) What is the **point estimate** of the population proportion **(P)?**

(Exam 2009)↓

(2) Develop a **95%** confidence interval for the **population proportion** and explain what it means.

(3) Does the sample evidence indicate that the sub-contractor is not maintaining the quality of his supplies at the agreed level? Use the significance level of 0.01.

(4) What is the minimum sample size required to estimate the proportion in Part (1) to within 0.03 at the 95% confidence level if:

(a) The sample proportion **(P)** is used as an estimate for the population proportion **(P)?**

(b) No previous estimate of **P** is available**?**

Q42. **Al Ahly** and **Al Zamalek** are the most famous football clubs in Egypt. **Al Ahly** supporters claim that their team is the best team in the Egyptian Football League. To investigate the validity of this claim, a sample of football matches was examined, and the number of attacking moves made per match was counted up. The mean and standard deviation of the number of attacking moves made per match were obtained for both teams. The results are given in the following table:

(Hypothetical Data)

Team	**Attacking Moves**		**Number of Matches**
	Mean	**Standard Deviation**	
Al-Ahly	$\bar{x}_1 = 30$	$s_1 = 5$	$n_1 = 50$
Al-Zamalek	$\bar{x}_1 = 24$	$s_2 = 8$	$n_2 = 50$

Let μ_1 is **the long** run mean number of attacking moves made by **Al Ahly** team,

μ_2 is **the long** run mean number of attacking moves made by **Al Zamalek** team. **(Exam 2009)** ↓

(1) Make a **99%** confidence interval for the **difference** between the two population means $(\mu_2 - \mu_1)$.

(2) What is the **maximum error** for the estimate of **Part (1)?**

(3) Based on the results of Part (1), can you conclude that the two teams are of the same level? Explain.

(4) Does the **first** sample provide sufficient evidence to conclude that the long run mean number of attacking moves made by **Al Ahly** is greater than **27?** Use α **= 0.05**.

(5) Using the level of significance **0.01,** evaluate the claim that **Al Ahly** team is **much better** than **Al zamalek** team.

(Exam 2009)

Q43. **Al-Ahly** and **Al-Zamalek** are the most famous football clubs in Egypt. **Al-Ahly** supporters claim that their team is the **best** team in the Egyptian Football League. To investigate the validity of this claim, a sample of football matches was examined, and the number of attacking moves per match was counted up. The **mean** and **standard deviation** of the number of attacking moves made per match were obtained for both teams. Following are the results:

(Hypothetical Data)

Team	Attacking Moves		Number of Matches
	Mean	Standard Deviation	
Al-Ahly	$\bar{x}_1 = 20$	$s_1 = 4$	$n_1 = 36$
Al-Zamalek	$\bar{x}_2 = 16$	$s_2 = 6$	$n_2 = 36$

Let μ_1 is the **long run mean** number of attacking moves made by **Al-Ahly** team,

μ_2 is the **long run mean** number of attacking moves made by **Al-Zamalek** team.

(Exam 2013) ↓

(1) Determine a **95%** confidence interval for μ_1. What is the **maximum error?**

(2) What **size** sample would have to be taken in order to **reduce** the **width** of the confidence interval obtained in **Part (1)** by **0.5?**

Hint: Other conditions remain the same as in **Part (1)**.

(3) Make a **99%** confidence interval for the **difference** between the two population means $(\mu_1 - \mu_2)$.

(4) Based on the results of **Part (3)**, can you conclude that the two teams are of the same level**? Explain**.

(5) Comment on whether it would be possible to reach a different conclusion for **Part (4)** using **95%** confidence level instead of **99%** confidence level.

Hint: No Calculations Required.

(6) Does the second sample provide sufficient evidence to conclude that the long run **mean** number of attacking moves made by **Al-Zamalek** team is **less than 18?** Use $\alpha = 0.01$.

(7) Using the level of significance **0.05**, evaluate the claim that **Al-Ahly** team is **much better** than **Al-Zamalek** team.

Q44. The operations manager of a large production plant would like to estimate the **average** amount of time a worker takes to assemble a new electronic component. **How large a sample** of workers should he take if he does not want the error of his estimate to exceed **24 seconds** using a **95%** confidence level**?** If**:**

(1) From an earlier study, the manager knows that the **standard deviation** of amount of time taken by workers to assemble a similar device is **2** minutes.

(2) After observing a number of workers assembling similar

(Exam 2013) ↓

devices, he noted that the **shortest** time taken was **8** minutes, while the longest time taken was **18** minutes.

Q45. The management of a supermarket wanted to investigate if the **percentages** of **men** and **women** who prefer to buy national brand products over the store brand products are **different**. A sample of **400 men** shoppers at the company's supermarkets showed that **160** of them prefer to buy national brand products over the store brand products. Another sample of **250 women** shoppers at the company's supermarkets showed that **90** of them prefer to buy national brand products over the store brand products.

Let $\mathbf{P_1}$ is the proportion of **all** men shoppers who prefer to buy national brand products,

$\mathbf{P_2}$ is the proportion of **all** women shoppers who prefer to buy national brand products.

(1) What is the **point estimate** of the **difference** between the two population proportions $\mathbf{(P_2 - P_1)?}$

(2) Construct a **95%** confidence interval for the **difference** between the proportions of **all** men and **all** women shoppers $\mathbf{(P_1 - P_2)}$ at these supermarkets who prefer to buy national brand products over the store brand products.

(3) Testing at the **5%** significance level, can you conclude that the proportions of all men and all women shoppers at these supermarkets who prefer to buy national brand products over the store brand products are **different?**

(4) Based on the results obtained in **Part (2)**, how to **verify** your results in **Part (3)?**

(Exam 2013) ↓

Q46. A random sample of **400** fish is taken from a lake, marked, and returned to the lake. After a suitable interval, a second sample of **400** is taken and **16** of these are found to be marked.

(1) Find the **point estimate** of the **proportion** of **marked fish** in the lake.

(2) Estimate the **number of fish** in the lake.

(3) Obtain a **95%** confidence interval for the **number of fish** in the lake.

(Exam 2013)

Q47. A large public utility company wants to compare the consumption of electricity during the summer season for a single - family houses in two cities that it services. For each household sampled, the monthly electric bill ($) is recorded with the following results:

City	Mean	Standard Deviation	Number of Households
City (1)	$\bar{x}_1 = 50$	$s_1 = 18$	$n_1 = 400$
City (2)	$\bar{x}_2 = 48$	$s_2 = 12$	$n_2 = 324$

Let μ_1 is the population **mean** monthly electric bill for **city (1)**,
μ_2 is the population **mean** monthly electric bill for **city (2)**.

(1) Construct a **99%** confidence interval for μ_2. What is the **maximum error?**

(2) What sample **size** would have to be taken in order to **reduce** the **width** of the confidence interval obtained in **Part (1) to 2.5?**

Hint: Other conditions remain the same as in **Part (1)**.

(3) Set up a **95%** confidence interval for the **difference** between the two population means **$(\mu_2 - \mu_1)$**.

(4) On the basis of your results of **Part (3)**, can you conclude

(Exam 2014) ↓

that the two population means **(μ_1** and **μ_2)** are **equal? Explain**.

(5) Using the significance level **$\alpha = 0.05$**, is there evidence that the mean monthly bill in **city (1)** is **above $47?**

(6) Is there evidence that the mean monthly bill is **lower** in **city (2)** than in **city (1)**? Use the level of significance **0.01**.

(7) Suppose that the utility company wants to **estimate** the population **mean** monthly electric bill for **city (1)**, **μ_1**, to **within ± 1.5** dollars with **95%** confidence. Because it does not have access to previous data, it makes its own independent estimate of the standard deviation, which it believes to be **15** dollars. **How large** a **sample size** would be required?

Q48. A public opinion organization wants to estimate the proportion of voters who will vote for the candidate (A) in a presidential campaign. In a survey of **500** voters, **70%** indicated that they will vote for this candidate.

(1) Construct a **99%** confidence interval for the **population proportion (P)**.

(2) At the **0.05** level of significance, can you conclude that the **population proportion (P)** is **greater than 65%**.

(3) What **sample size** is needed if the organization wants to be **95%** confident of being correct to within **±0.03** of the true population proportion **(P)**? **Hint:** Use the sample proportion as an estimate for the population proportion **(P)**.

(4) If the organization has not previously undertaken such a survey, find the **conservative** sample size for **Part (3)**. Comment on your results of **Parts (3)** and **(4)**.

(Exam 2014)

Q49. **Two** insect sprays are to be compared. **Two** rooms of equal size are sprayed, one with **spray (1)** and the other with **spray (2)**. Then **100** insects are released in each room, and after 2 hours the Dead insects are counted. Suppose the result is **64** dead insects in
the room sprayed with **spray (1)** and **52** dead insects in the other room.

Let $\mathbf{P_1}$ is the **proportion** of **dead insects** in the room sprayed with **spray (1)**,

$\mathbf{P_2}$ is the **proportion** of **dead insects** in the room sprayed with **spray (2)**.

(1) Make a **95%** confidence interval for the **difference** between the two population proportions $\mathbf{(P_2 - P_1)}$.

(2) Based on the results of **Part (1)**, can you conclude that the **two sprays** are of the **same effect** for killing insects**? Explain**.

(3) Comment on whether it would be possible to reach a different conclusion for **Part (2)** using **99%** confidence level instead of **95%** confidence level.

Hint: <u>No Calculations Required</u>.

(4) Are these data strong enough to conclude, at the **0.05** significance level, that **spray (1)** is **more effective** than **spray (2)** for killing insects?

Q50. A market researcher for a consumer electronics company wants to study the television viewing habits of residents of a particular small city. A random sample of **64** respondents is selected, and each respondent is instructed to keep a detailed record of all television viewing in a particular week.
The results are as follows**:**

- **Viewing time per week:** $\bar{x}$ **= 20** hours, **S = 4** hours.
- **48** respondents **watch the evening news**.

(Exam 2015) ↓

Let **μ** is the **mean** amount of television watched per week in this city,

P is the **proportion** of respondents who watch the evening news.

(1) Set up a **99% confidence interval** estimate for the **mean** amount of television watched per week in this city (**μ**). What is the **maximum error?**

(2) What **size** sample would have to be taken in order to **reduce** the width of the confidence interval obtained in **Part (1)** by **0.5?**

Hint: Other conditions remain the same as in **Part (1)**.

(3) On the basis of your answer to **Part (2)**, what general conclusion can be reached about the **effect** of the acceptable **sampling error** on the **sample size** needed? **Discuss**.

(4) What is the **point estimate** for the **population proportion (P)**?

(5) Make a **95%** confidence interval for the **proportion** of respondents who watch the evening news **(P)**.

(6) At the **0.01** level of significance, is there evidence to believe that the true **mean** amount of television watched per week (**μ**) in this city is **less than 22**?

(7) It is claimed that the **proportion** of respondents who watch the evening news **(P) differs** from **0.6**. Investigate the **validity** of this **claim** using **α = 0.05**.

(8) If the market researcher wants to take another survey in a different city, **How large** a sample is needed if he wishes to be **95%** confident that his sample proportion will not differ from the true proportion by more than **0.035**?

Hint: No information available about the **population proportion**.

(Exam 2015)

Q51. A marketing manager for a company wishes to **compare** the **mean prices** charged for **two brands** of a product. The manager conducts a random survey of retail outlets and obtain independent random samples of prices with the following results**:**

	Brand (A)	**Brand (B)**
Sample **mean ($\bar{x}$)**	50	46
Sample **standard deviation (s)**	12	10
Sample **size (n)**	100	125

Let μ_A is the **mean** price for **Brand (A)** and μ_B is the **mean** price for **Brand (B)**.

(1) Find a **95% confidence interval** to estimate the **mean** price of **Brand (B)**.

(2) Can the manager conclude that the **mean** price of **Brand (B) differs** from **45**? Use the **0.05** significance level.

(3) **Comment** on your results of **Part (2)** as **compared** with that of **Part (1)**.

(4) Suppose the **confidence interval** obtained in **Part (1)** is **too wide**, and the manager wishes to **reduce** the width of this confidence interval **by 0.5**, how **large** a sample is required?

Hint: Other conditions remain the same as in Part (1), and the sample standard deviation (s_B) is used as an estimator for the population standard deviation (σ_B).

(5) What is the **point estimate** for the **difference** between the two population means **(μ_B - μ_A)**.

(6) Calculate a **99%** confidence interval for **(μ_A - μ_B).** Can we be **99%** confident that **μ_1** and **μ_2 differ**? **Explain**.

(7) Use an appropriate hypothesis test to provide evidence

(Exam 2016) ↓

supporting the **claim** that the **mean** price of **Brand (B)** is **lower than** the **mean** price for **Brand (A)**. Use **α** = **0.01**.

Q52. A social researcher wants to determine if people think that there is too much violence on television these days. **60%** of **200 men** selected at random believed so. **80%** of **250 women** selected at random also believed that there is too much violence on television.

Let $\mathbf{P_1}$ and $\mathbf{P_2}$ be the **proportions** for **all men** and **all women**, respectively.

(1) Find a **99%** confidence interval for the **difference** between the two population proportions $\mathbf{(P_1 - P_2)}$.

(2) Carry out the appropriate **test** to investigate if there sufficient evidence to conclude if there is a **significant difference** in the **views** of **men** as **compared** to views expressed by **women**? Use **α = 0.05**.

(3) **(3)** How would your conclusion of **Part (2) change** if the level of significance is changed to **0.01**?

(Exam 2016)

Q53. To determine the effectiveness of a new method of teaching reading to young children, a group of 100 nonreading children were randomly **divided** into **two** groups of **50** each. The **first** group was taught by a **standard** method and the **second** group by an **experimental** method. At the end of the school term, a reading examination was given to each of the students, with the following summary statistics resulting**:**

	Standard	**Experimental**
Average Score ($\bar{x}$)	65	70
Standard deviation (s)	5	4

(Exam 2017) ↓

Let μ_s be the **mean** score for **the standard method** and
μ_e is the **mean** score for the **experimental** method.

(1) Develop a **99% confidence interval** to estimate the **mean** score for the **standard method (μ_s)**.

(2) Assuming that the sample standard deviation and the confidence level remain the same as in **Part (1)**, **how large** should the sample have been to **increase** the **width** of the confidence interval found in **Part (1) to 4**?

(3) How can you **explain** the **difference** between the **two sizes** of the sample as far as the **maximum error** is concerned?

(4) What is the **point estimate** for the population mean **(μ_e)**.

(5) Find a **95% confidence interval** for **($\mu_e - \mu_s$)**.

(6) Are these data strong enough to provide, at **α = 0.01**, that the **experimental** method results in a **higher** mean test score than that of the **standard method**?

Q54. In a public opinion poll, suppose that in a sample of **200** people from **City (A)**, **70%** favored a certain policy, and in a sample of **250** people from **City (B) 60%** favored this policy.
Let P_A be the **proportion** of people who **favored** the policy in **City (A)**, and P_B be the **proportion** of people who **favored** the policy in **City (B)**.

(1) Find a **95% confidence interval** for the population proportion **(P_B)**.

(2) Does the sample from **City (A)** provide sufficient evidence to conclude, at **α = 0.05,** that the proportion of people who favored the policy **(P_A)** is **different from 0.8**?

(3) It is claimed that this policy is **less** favored in **City (B)** than in **City (A)**. Investigate the **validity** of this claim using the significance level **0.05**.

(Exam 2017) ↓

(4) For a third **city (C)**, **how many** people should be sampled to be **95%** confident that the sample proportion **will not differ** from the true proportion **(P_c)** by more than **0.02**?
Hint: No information available about **(P_c)**.

(Exam 2017)

Q55. A sales manager wishes to compare the effectiveness of two methods for training new salespeople. He selects 200 sales trainees who are randomly divided into **two** experimental groups of **100** each. The salespeople are then assigned and managed without regard to the training they have received. At the year's end, the manager reviews the performances of salespeople in these groups and finds the following results**:**

	Group A	**Group B**
Average Daily Sales (x)	**$150**	**$145**
Standard Deviation (s)	**20**	**25**

Let **μ_A** and **μ_B** be the **mean** daily sales for **type A** training and **type B** training, respectively.

(1) Calculate a **95% confidence interval** for the **mean** daily sales of **type A** training **(μ_A)**.

(2) What **sample size** needed in order to **reduce** the **width** of the confidence interval obtained in **Part (1) by 1.5**?
Hint: Other **conditions** remain **the same** as in **Part (1)**.
What is the **change** that happened to the **sample size**? **Explain**.

(3) Does the sample of type B provide sufficient evidence to conclude that the population mean (μB) is less than 150?
Use α = 0.05. **(Exam 2018)** ↓

(4) What is the **point estimate** for the **difference** between the mean daily sales **($\mu_B - \mu_A$)**.

(5) Find a **99% confidence interval** for **($\mu_B - \mu_A$)**. What would you **conclude** about the **equality** or **inequality** of the **two means**? **Explain**.

(6) At a level of significance **0.01,** is there evidence that **type A** training produces results that are **superior** to those of **type B**?

Q56. Assume that independent simple random samples of tax returns from two offices provide the following information**:**

Office (1)	Office (2)
$n_1 = 250$ Number of returns with errors = 35	$n_2 = 300$ Number of returns with errors = 30

Let P_1 and P_2 be the **proportion** of returns with errors from **Office (1)** and **Office (2)**, respectively.

(1) What is the **point estimate** of the population proportion **(P_1)**?

(2) It is claimed that the **proportion** of tax returns from **Office (1) exceeds 0.1**? **Investigate** the **validity** of this claim using **$\alpha = 0.05$**.

(3) Develop a **95% confidence interval** for the difference between the two population proportions **($P_1 - P_2$).**

(4) Do the samples provide sufficient evidence to conclude that there is a **significant difference** between the two population error rates P_1 and P_2? Use the significance level **0.01**.

(Exam 2018)

Q57. Suppose that we **reject** a **null hypothesis** at the **5%** significance level, for which of the following levels of significance do we **also reject** the null hypothesis?

(A) 4% **(B) 2.5%** **(C) 6%** **(D) 3%**

Q58. Which of the following is **not a correct** way to state the **null hypothesis**?

(A) $H_0: \widehat{P}_{11} - \widehat{P}_{12} = 0$ **(B)** $H_0: \mu = 10$

(C) $H_0: \mu_1 - \mu_2 = 0$ **(D)** $H_0: P = 0.5$

Q59. The statement "if there is sufficient evidence to reject a null hypothesis at the 5% significance level, then there is sufficient evidence to reject it at the 1% significance level" is

(A) Always true **(B) Never true**

(C) Sometimes true; the value of the test statistic needs to be provided for a conclusion

(D) Not enough information; this would depend on the type of statistical test used

Q60. All of the following **increase** the **width** of a **confidence interval except**

(A) Increased confidence level **(B) Increased variability**

(C) Increased sample size **(D) Decreased sample size**

Sample weights (in pounds) of newborn babies born in two **countries (A and B)** yielded the following data**:**

Country (A): Sample **Size** = **75** , Sample **Mean** = **6.12**, Sample **Variance** = **9**

Country (B): Sample **Size** = **100** , Sample **Mean** = **7.26**, Sample **Variance** = **13**

Let μ_A and μ_B are the **mean** weight for the two countries **A** and **B**, respectively.

Answer the following <u>Three</u> Questions (Q61- Q63):

(Exam 2019) ↓

Q61. The **upper limit** of the **95% confidence interval for** $(\mu_B - \mu_A)$ is

(A) 2.18 (B) 1.94 (C) 2.12 (D) 1.88

Q62. To investigate the validity of a researcher's **claim** that the **mean** weight of newborns in **City (B)** is **greater than** that in **City (A)**, the following test was performed at the **1%** significance level**:**

H_0**:** $\mu_B - \mu_A = 0$ versus H_1**:** $\mu_B - \mu_A > 0$.

For this test, the **value** of the **test statistic** is

(A) 1.96 (B) 2.28 (C) 2.42 (D) 1.86

Q63. Refer to Q33, which of the following is the **most appropriate conclusion**?

(A) Reject the null hypothesis; there is sufficient evidence to support the research's claim.

(B) The null hypothesis is rejected; there is no sufficient evidence to support the researcher's claim.

(C) The null hypothesis is not rejected; there is sufficient evidence to support the researcher's claim.

(D) Fail to reject the null hypothesis; there is no sufficient evidence to support the researcher's claim.

Q64. Determine the **minimum** required **sample size** if you want to be **99%** confident that the sample mean is **within 3** units of the population mean given $\sigma = 8$.

(A) 48 (B) 47 (C) 46 (D) 50

Q65. A **confidence interval** for a **population proportion** is **0.41775** and **0.58225**. If you know that the **sample size** is **100**,

(Exam 2019) ↓

determine the **confidence level** that can be attached to this interval.

(A) 95% (B) 98% (C) 90% (D) 99%

Hint: For computations, no rounding required.

Q66. A statistical report states that the **95% confidence interval** for the **mean** score of students is between **20** and **25** obtained using a **sample** of **144** students. Determine the **population standard deviation**.

(A) 15.3 (B) 15.6 (C) 16.5 (D) 16.2

A social researcher wants to determine if people think that there is too much violence on television these days. **70%** of **120 men** selected at random believed so. **50%** of **80 women** selected at random also believe that there is too much violence on television.

Let $\mathbf{P_1}$ and $\mathbf{P_2}$ be the **proportions** for all **men** and all **women**, respectively.

Answer the following three Questions (Q67 – Q69):

Q67. The **95% confidence interval** for the population proportion **(P_2)** is

(A) 0.1 and 0.9 (B) 0.3 and 0.7

(C) 0.2 and 0.8 (D) 0.4 and 0.6

Q68. To carry out a **test** for investigating if there is sufficient evidence to conclude if there is a **difference** in the views of **men** and **women** on this issue, the appropriate **null** and **alternative** hypotheses are**:**

(A) $H_0: P_1 - P_2 = 0$ vs $H_1: P_1 - P_2 > 0$

(B) $H_0: P_1 - P_2 = 0$ vs $H_1: P_1 - P_2 < 0$

(C) $H_0: P_1 - P_2 = 0$ vs $H_1: P_1 - P_2 \neq 0$

(D) $H_0: P_1 - P_2 = 0$ vs $H_1: P_1 - P_2 \geq 0$

(Exam 2019) ↓

Q69. Referring to Q39 and using **$\alpha = 0.01$**, the appropriate conclusion is

(A) The value of the test statistic is 2.15, and there is not a significant difference in the views of men and women.

(B) The value of the test statistic is 2.86, and there is a significant difference in the views of men and women.

(C) The value of the test statistic is 2.24, and there is not a significant difference in the views of men and women.

(D) The value of the test statistic is 3.2, and there is a significant difference in the views of men and women.

(Exam 2019)

References

1. Anderson, R. A., Sweeney, D. J. and Williams, T. A. (2002). Statistics for Business and Economics (8th edition). South - Western Thomson Learning.

2. Erichson, B. H. and Nosanchuk. T. A. (1977). Understanding Data. McGraw-Hill.

3. Lind, D. A., Marchal, W. G. and Wathen, S.A. (2008). Basic Statistics for Business and Economics (13th edition). McGraw-Hill.

4. Mann, P. S. (2007). Introductory Statistics (2007). John Wiley.